

# White Paper Report

Report ID: 100782

Application Number: HD-51084-10

Project Director: Julia Flanders (j.flanders@neu.edu)

Institution: Northeastern University

Reporting Period: 1/1/2011-12/31/2014

Report Due: 3/31/2015

Date Submitted: 7/16/2015

# **Final Performance Report**

**HD-51084-10**

## **A Journal-Driven Bibliography of Digital Humanities**

Project Director: Julia Flanders

Northeastern University

May 30, 2015

## Overview

This project began with a simple premise. Digital Humanities Quarterly is an online, open-access journal whose founding coincided with the founding of the Alliance of Digital Humanities Organizations (ADHO) in 2005, and whose topical scope covers all areas of the field we now know as “digital humanities.” The bibliographies of DHQ articles thus reflect the intellectual watershed of this field, and also its formation over the life of the journal itself. Under this grant we sought to aggregate these bibliographies into a central bibliographic database, with two goals. First, at a practical level we wanted to simplify the journal’s production workflow and eliminate the duplication of data resulting from storing bibliographic data in the articles themselves. With a centralized database, we could store authoritative bibliographic data in one place and reference it from the articles, taking advantage of the fact that many DHQ articles draw on a common pool of material for their citations. Second, from a research perspective this data clearly constituted a potential public good and a fascinating data set in its own right. With a centralized database, we would be able to study patterns of co-citation, learn about the evolution of the field, and study the citation practices of different subcommunities. Bibliographic data could also potentially serve as a way for readers to find articles of interest, or clusters of related articles.

We framed the effort as an 18-month process, with the project originally scheduled for completion in July 2012. Although this workplan was not unrealistic, retrospective analysis reveals its vulnerabilities: above all, because of the small size of the grant, we relied on commitments of donated effort for significant parts of the technical development work, notably the original data capture system and the integration of the new bibliographic data into the DHQ interface. As described in more detail below, one of the initial obstacles we faced was a set of problems with the data capture system which could not be addressed because the anticipated expertise was no longer available to us. Another more significant vulnerability was the fact that the data capture itself required fairly significant attention to issues of bibliographic genre and hence required a level of training and dedication that was somewhat out of proportion to the overall interestingness of the work, making it difficult to hire and retain students. As a result, there were periods of inactivity and delay while we searched for new research assistants. The third and most significant disruption could not have been predicted: in July 2013, the principal investigator changed jobs and moved from Brown University to Northeastern University, and DHQ moved its editorial operations to Northeastern at the same time. During the period of transition, work on this project was more or less suspended, and was not resumed until we hired a new research assistant in January 2014 who was able to bring the data capture and error correction to completion in December 2014 after three no-cost extensions.

This prolonged and constantly changing work process could look from some perspectives like a narrative of failure, and certainly there have been important lessons learned. However, this project also illustrates an important principle that informs the design of the DH Startup grant program, namely the fact that some kinds of work are especially unpredictable. Small-scale projects are more vulnerable to disruption because they tend to have fewer resources to fall back on, and because they are operating on small enough quantities of effort that even a small reduction makes a significant

difference. Because small-scale projects in academic settings often rely on student labor, they have the additional vulnerability that comes from unpredictable turnover. The ultimate successful outcome of this project owes a great deal to the flexibility afforded us by NEH, for which we are extremely grateful.

## **Project Activities**

### ***Main activities***

#### **Data Capture**

The initial capture of bibliographic data for this project was undertaken using a web-based bibliographic data capture and management system developed at the Brown University Center for Digital Scholarship for use in its digital humanities projects. The system offered a form-based data entry interface, with the data being saved as MODS. Configuration files permitted different projects to define different bibliographic genres and the required and permitted fields associated with each one, allowing a high degree of control which we felt was desirable for DHQ's purposes. Using this system, we established a set of bibliographic genres representing the requirements of DHQ's existing citations, and hired a group of undergraduate students to undertake the data capture. Our original goal as defined in the grant proposal was to capture bibliographic items not only from DHQ's own article bibliographies, but also items from the other major digital humanities journals (including *Computers and the Humanities* and *Literary and Linguistic Computing*), and we made significant progress on those two journals. However, changes to personnel and local support at Brown University interrupted that work process and we did not complete the capture of CHum and LLC data. We encountered two chief obstacles at this stage. First, the data capture system was engineered in a way that caused its performance to suffer dramatically under large quantities of data, and second, changes in personnel at Brown University reduced the levels of technical support available to us, so we were not able to address the problems with the data capture system, or add the features for de-duplication and error checking that we had anticipated. However, under this system we were able to capture a significant number of records (approximately 3000 in all).

After the move to Northeastern, we hired a graduate research assistant to complete the data capture, and we also faced the fact that we needed to adopt a different data capture tool and process. Although the web data entry interface of the Brown tool had significant advantages of ease of use, our new graduate assistant had greater familiarity with XML and we anticipated that once the data capture was complete our general DHQ workflow would rely on DHQ's managing editors (also comfortable with XML), so a form-based system would not be necessary. In addition, the remaining data capture was focused on the bibliographies of existing DHQ articles which were already expressed as lightly encoded XML, so we could benefit from using XML tools to convert them into our target format. At this stage we developed a schema (described in more detail below) that reflected the genres of bibliographic record we had already established (including their requirements for the presence and order of fields) and set up a work flow to convert these bibliographies. The first step in the process involved an XSLT stylesheet that converted the existing TEI elements into the corresponding bibliographic elements

of our schema, wrapped in a generic <BiblioItem> element. The second step involved hand editing these records to change the wrapper element to a more specific one reflecting the genre of the item (e.g. <Book>, <BookSection>, <JournalArticle>, etc.) and to add further detailed markup of the individual components of the entry that were not available in the original DHQ encoding. (Because that encoding was driven by display needs rather than by goals of bibliographic completeness, only titles and URLs were typically explicit in that markup.)

Following the completion of the data capture, there was some further work involved in cleaning up the data:

- Some de-duplication was necessary, since the initial data capture had been done in a system that did not make it easy to check for the existence of a given record before entering it.
- We had to ensure that record IDs were unique. IDs for bibliographic items in the system were based on author and date rather than on randomly assigned identifiers, to make it easier to spot errors of citation in the encoding of DHQ articles, but the author-date system requires disambiguation for common surnames and for authors who publish multiple items in a single year.

As part of the cleanup process we also had to consider and document our policies concerning the level of bibliographic management we were prepared to exercise. For example, in cases where different DHQ articles cited different versions of the same published item (for instance, hardcover and the paperback editions, published in different years), we decided to treat these as separate items rather than develop a mechanism for coordinating them; at a later stage we may institute a formal mechanism for representing these connections in the data to improve analysis. Similarly, we do not track connections between versions of published items (such as a blog post that is republished in a journal and then anthologized in a book). We also determined that some kinds of cited items did not belong in the centralized bibliography at all, the primary example being items that had only local relevance within the context of a specific DHQ article, such as personal communications (“Private email to the author, 31 May 2010” and the like). These items would remain in the separate DHQ articles and would not be aggregated centrally.

### **Bibliographic Identifiers in DHQ articles**

Once the bulk of the data capture was complete, the next step was to establish the linkage between DHQ articles and the bibliographic items they cite. All published DHQ articles include full bibliographies, and in our earlier practice any citations in the text pointed to entries in those bibliographies, as in the following example:

Inline reference in the body of the article:

```
<ptr target="#mcgann2004" loc="50–1" />
```

Bibliography entry:

```

<bibl xml:id="mcgann2004" label="McGann
2004"><author>McGann, J. J.</author> <title
rend="italic">Radiant Textuality: Literature After the
World Wide Web</title>. <pubPlace>New York</pubPlace>:
<publisher>Palgrave</publisher>,
<date>2004</date>.</bibl>

```

The @target attribute of the <ptr> element is a local URL that points to the @xml:id attribute of the <bibl> element, establishing a link between them. When the article is published, an XSLT stylesheet finds each <ptr> element, follows the link and takes the value of the @label attribute to be used in the display as a link to the appropriate bibliography entry. The entry itself is transformed by the stylesheet as well to display according to the journal's standard format:

McGann 2004. McGann, J. J. *Radiant Textuality: Literature After the World Wide Web*. New York: Palgrave, 2004.

In establishing the new system, we needed to consider both the desired endpoint of the process (a working publication in which all bibliographic data would be centralized) and also the intermediate steps, which included the need to verify the accuracy of links to the centralized bibliography, and also the need to provide a fallback in case of broken or missing data. We did not want to throw away the bibliographic data we already had in place until the very end of the process (if then). The process we followed was:

1. Create a second attribute for <bibl> that would carry a pointer to the centralized bibliography, and populate it with provisional values, using the existing values of @xml:id. Since these values were based on the author and date of the item, we reasoned that those would often correctly identify the intended item in the centralized bibliography. We created a new @key attribute and globally propagated the existing value of @xml:id to @key. The existing internal pointers that link the inline references to the article's bibliography are left in place unchanged.
2. Check for non-existent records (that is, cases where the @key value does not match any existing record in the centralized bibliography) and for incorrect links (that is, cases where the value of @key points to the wrong entry in the centralized bibliography). For this purpose we created an XSLT stylesheet that took each article's bibliography, and for each item used its @key value to identify and pull in the matching record (if any) from the centralized bibliography. The stylesheet displayed this information in tabular form with the original entry and the matching entry side by side for comparison. It also performed a comparison of the title fields in the two entries to determine whether they were likely to represent the same bibliographic item, and it looked as well for other entries with similar titles which might be alternative matches (or possible duplicate records). Finally, it generated a color-coded border identifying probable errors: red for cases where no matching entry was found, yellow for cases where the title match was questionable, and green for entries that matched both the @key and the title

similarity test. Using this display, we reviewed all of the published DHQ articles, added missing entries, fixed errors, and resolved ambiguities. For purely local references (the “private email to author” case given above), we added a @key=“[unlisted]” on the <bibl> to signal that no link to the centralized bibliography was needed.

3. Provide authors with a similar side-by-side view of the bibliography for their article, so that they have an opportunity to verify the accuracy of the data. This precaution serves as a fallback in case of oversight during what were necessarily quite repetitive and large-scale tasks (and hence prone to occasional slips). This process was not completed under the grant, but is now being undertaken by DHQ in summer 2015.
4. Update the DHQ display stylesheets so that instead of using the local bibliography for each article, they draw data from the centralized bibliography. As part of this process, we also had to develop new display logic to use the fully encoded data from the centralized database (which does not include literal punctuation such as periods, commas, quotation marks, etc. to delimit the individual fields). These updates have been completed and are awaiting the completion of the author check before we switch over to using the centralized data. We anticipate that we will be using the new system starting in fall 2015.
5. Discard the original bibliographic data? In theory, once we have been using the centralized bibliography for long enough to be comfortable that it is complete and accurate, we will have no further need for the locally encoded bibliographic data. Because the entire system is maintained under version control, we can delete this information without truly losing it, in case we need to check it or retrieve it at some future point.

The final encoding looks like this:

Inline reference in the body of the article:

```
<ptr target="#mcgann2004" loc="50–1" />
```

Local bibliography entry in the article:

```
<bibl xml:id="mcgann2004" label="McGann 2004"
key="mcgann2004a"><author>McGann, J. J.</author> <title
rend="italic">Radiant Textuality: Literature After the
World Wide Web</title>. <pubPlace>New York</pubPlace>:
<publisher>Palgrave</publisher>,
<date>2004</date>.</bibl>
```

Remote entry in the centralized bibliography:

```
<Book ID="mcgann2001a" issuance="monographic">
  <author>
    <givenName>Jerome</givenName>
    <familyName>McGann</familyName>
```

```

        </author>
        <title>Radiant Textuality: Literature After the World
Wide Web</title>
        <place>New York</place>
        <publisher>Palgrave Macmillan</publisher>
        <date>2001</date>
    </Book>

```

Note that the internal linking between <ptr> and <bibl>, and the generation of a display label, is left untouched and is purely local to the article; the disambiguation of entries required in the centralized resource (e.g. “mcgann2004a”, “mcgann2004b”, etc.) is not necessary or visible within the article itself unless the article itself references more than one 2004 item for McGann. This separation of local and external ecologies had the added benefit of avoiding the necessity of updating the @target and @xml:id values, which would have added significant work and opportunities for error.

### Design of publication system

The bibliographic data resource developed under this grant represents a new level of complexity for the DHQ publication, since it exists as a separate data set referenced from the DHQ articles, and the publication process needs to follow the bibliographic pointers from the articles to retrieve the relevant bibliographic records and incorporate them appropriately into the article’s display. Additionally, the existence of the bibliography as a distinct resource opens up possibilities for analysis of this resource in its own right. Both of these things can be accomplished using our existing architecture: XSLT stylesheets for the transformation of data from TEI into HTML, and the Apache Cocoon pipelining system to provide the overall user interaction logic, navigation, and site organization. However, the more natural tool to use as DHQ gains in complexity is an XML database through which the data could be indexed, searched, and processed more efficiently. We are currently exploring the use of eXist (an open-source XML database) as a next step for this project, but this carries some overhead of development and maintenance that lies outside the immediate scope of this project.

### Visualization experiments

The final component of this project was the analysis and visualization of the bibliographic data, which was done in partnership with two groups at Indiana University. Our original plan included a collaboration Katy Börner’s research team at the Center for Network Science, and at intervals during the project we provided preliminary data sets for experimentation. Based on early discussions with the visualization team we developed a specification for exporting the combined DHQ article and bibliographic data in a spreadsheet format that supported the types of analysis we were most interested in: comparisons of DHQ articles based on co-citation, with DHQ article metadata (chiefly author affiliations and abstract) as additional facets of analysis. Later in the process, once the data capture and cleanup were close to complete, we provided a fuller data set to Scott Weingart (a member of Börner’s research group) who performed some initial analysis. Following the conclusion of the grant, we will continue to work with Weingart to take the analysis further. Because of the challenges



encountered earlier in the project, we did not get as far with the visualization work as we had initially hoped, but we did accomplish all of the parts that required active funding support; the foundation we have established under this grant will enable us to proceed with DHQ's own resources.

Fortuitously, we were also able to undertake a second collaboration on visualization of bibliographic data which though not formally part of this grant project is very closely tied to it. Immediately following the conclusion of the grant, in the spring semester 2015, DHQ participated as a client project in the Information Visualization MOOC offered at Indiana University, making our data available to a team of student researchers as the basis for a research project in visualization. The students developed a set of visualizations and a detailed analysis of citation patterns, and provided an extensive final report. Members of the DHQ editorial team will be collaborating with the student team to produce a co-authored article based on this report, to be published in DHQ later this year, together with the resulting visualizations. Samples are included in the appendix to this report.

### ***Reasons for changes and omissions***

As noted in the introduction to this report, this project deviated significantly from its original work plan. There were some modifications to the timing and duration of activities that resulted from institutional changes over which DHQ had no control: changes to staffing and level of technical support at Brown University, and the 2013 move of DHQ's editorial operations to Northeastern as a result of Julia Flanders' institutional move. There were also some modifications to the overall scope of the project. In our original work plan we had planned to work with arts-humanities.net (which at that time was managing a bibliographic tool as well) on shared management of bibliographic records, but arts-humanities.net ceased operations shortly after the start of this project and that collaboration was not possible. At a future time it may prove possible to host a contributory interface for DH bibliography, perhaps hosted through the Alliance of Digital Humanities Organizations, but that would need to be a community decision supported by community funding. In our original proposal we had also planned to include complete coverage of materials published in other DH journals (including Vectors, LLC, Digital Studies/Le Champ Numérique, and Text Technology) but the process of data capture proved more labor-intensive than we had expected and the data capture system itself did not mature technologically as we had planned (lacking anticipated support from Brown), so that processes like de-duplication were not as efficiently accommodated. At a future time we hope to have opportunities to ingest and integrate these other bibliographies, particularly if there turns out to be community support for a comprehensive bibliography of DH.

### ***Changes in Methods Involving Technology***

As noted in an earlier report, our original data capture system proved to have significant weaknesses. It was good at profiling data in an appropriately detailed manner, but it proved too slow for efficient use. As part of this grant, we did an extensive data profiling exercise and developed a schema that matches the MODS profile used internally within the original data capture system, but provides better constraint based

on specific bibliographic genres. MODS was appropriate within a web-based data capture environment, since all of the relevant constraint in that case was provided by the web form itself. However, in our new capture environment (using the Oxygen XML editor and relying on the schema to provide constraints), we needed a schema that would, for instance, stipulate that “book” items required a publisher field, whereas “blog post” items would not. The MODS schema is too permissive to provide such constraints, and it also provides very little precision in the semantics of specific elements. (For instance, a journal title is represented using a <mods:title> element within a <mods:relatedItem> element.) The data capture schema we developed provides a much simpler and more direct set of constraints for specific bibliographic genres such as books, book chapters, journal articles, conference papers, art works, blog posts, web pages, white papers, and other common forms of publication. For each genre, we identified the bibliographic elements that would be required and permitted, enabling us to establish consistency and test for missing required components.

It is worth noting that this schema is intended for internal purposes, and is not intended as a quixotic attempt to create yet another perfect bibliographic data format. Our goals in modeling this data are:

- To provide the constraint necessary to ensure consistency of data
- To provide enough semantic explicitness to permit mapping the data onto other bibliographic formats (such as TEI, MODS, etc.)
- To provide enough granularity to support the necessary display logic so that individual entries could be punctuated and formatted appropriately within the context of the DHQ publication interface

In other words, we do not expect other projects to use this schema, but we do expect that we will be able to map bibliographic data in other formats onto this one when we want to ingest data from other sources, and we also expect to be able to export data from this format into other formats as needed.

For the new data capture, we are using the Oxygen XML editor. We set up a “project” in Oxygen that permits validation, uniqueness checking, and XSLT transformations across the entire data set (which is broken up into multiple files to reduce lag). As new items are added, the system automatically runs a comparison across the data set to check for items with similar authors and titles (so as to flag potential duplicates). It also checks the uniqueness of the author-title identifier that serves as the unique key for individual entries within the system. Finally, using XSLT and CSS we can provide a basic visual display of the data when needed, e.g. for proofreading.

### *Efforts to publicize*

We have publicized our goals and progress for this project at several points. An export of our journal and bibliographic data was shared with the Information Visualization MOOC held at Indiana University in 2014-15, and served as a client project for a student working group in that course. An article reporting on their analysis will be published in DHQ later in 2015. Regular reports on progress have been included in DHQ’s annual reports to the Alliance of Digital Humanities Organizations. A presentation on the

project was made at the DH2015 conference in Sydney, Australia in July 2015. Once we complete the final integration of the bibliographic data into DHQ's publication interface, we will announce the completion of the project and its outcomes in a posting to the Humanist listserv, as well as via DHQ's regular dissemination mechanisms (including Twitter and the DHQ web site).

## Accomplishments

The accomplishments resulting from this project are as follows:

1. We digitized over 6000 bibliographic items covering all items referenced by DHQ articles, plus incomplete but substantial coverage of bibliographies from articles published in *Computers and the Humanities* and *Literary and Linguistic Computing*. Our original goal was to capture all bibliographies from CHum and LLC, plus conference proceedings from the DH conferences, but we were unable to get this data in a form we could easily convert and import, and it was not practical to capture it or convert it by hand.
2. We developed a schema for DHQ's bibliographic data, which is fine-grained enough to support export into other bibliographic formats (such as MODS or TEI).
3. We developed a set of additional tests and quality assurance mechanisms using Schematron and XSLT that support de-duplication and data integrity checking as part of DHQ's regular publication work flow.
4. We developed display stylesheets to support the integration of centralized bibliographic data into the DHQ publication interface.
5. In partnership with researchers at Indiana University (both within the Center for Network Science and through the Information Visualization MOOC), we developed visualizations that exploit the DHQ article metadata and bibliographic data.

Following the completion of the grant, we plan the following additional work:

1. Continue to expand the centralized bibliography as new DHQ articles are published; resources permitting, expand the bibliography by ingesting or capturing additional records (e.g. from the DH Conference Abstracts database, or from other journals).
2. Develop further visualizations as we expand our metadata. For instance, we are now working on adding topical keywords to DHQ articles, and these would support visualizations showing the citation patterns of articles on specific topics.
3. Integrate a dynamic bibliographic visualization into the DHQ web site. This will require that we serve the bibliographic data dynamically from an XML database, so that users can interact with it.
4. Make the bibliographic data available for public download so that others can experiment with it; eventually, we plan to develop an API to the bibliographic data to facilitate experimentation.

5. Develop an interface to the bibliography itself, so that readers can search, sort, and view items and learn more about citation and publication practices in digital humanities. As the field continues to develop, this bibliography will become an important instrument for studying the history of the field through its publications.
6. Implement authority control for the major informational components of these records (such as author names, publishers, and locations) to enhance consistency and ease data entry.

## **Audiences**

One primary audience for this work is DHQ's existing readership, who will receive the bibliographic data seamlessly integrated into the DHQ interface. These readers will benefit from greater consistency in formatting and presentation of the data, and also from greater accuracy in the citations (since authors often omit or misstate specific pieces of bibliographic information and these errors are not always caught prior to publication).

Another related audience is the members of the DH community who are interested in learning about the DH field through its patterns of citation and publication practices. This audience will be able to get a more detailed view of the field through the ability to query and analyze the bibliography. As the bibliography continues to grow, this audience will have an increasingly rich resource to work with. Providing the data for download and via an API will serve the smaller sector of this audience who are interested in doing their own data analysis.

Finally, an important "audience" for this work is DHQ's own internal community, especially including our production team. One primary motivation for this project was to eliminate duplication of data and to implement a more streamlined, data-driven approach to the bibliographic aspects of our publication. While this new system will not hugely reduce the overall work involved, it will shift the emphasis of that work from tasks that are annoying and demoralizing (i.e. copyediting of bibliographic minutiae) to tasks that contribute to the growth of knowledge in the field (i.e. enhancing the bibliographic data itself).

## **Evaluation**

As the introductory section of this report illustrates, the design and planning of this project contained several significant weaknesses, most notably an over-reliance on a tool for which we could not take technical responsibility. It also suffered from a lack of strong project management as a result of the fact that the principal investigator was overseeing several other grant-funded initiatives and other projects. These are both classic difficulties for digital humanities projects, but knowing about these risks in advance would not necessarily have enabled us to avoid them; the reason we chose to use Brown's bibliographic tool was its ease of use, proximity, and fitness for purpose; alternatives we considered all would have been either more expensive (i.e. out of scope for the project) or much less well adapted for the work. And at the time that we submitted the application, the other grants that competed for the principal investigator's

attention had not been awarded. On balance, we made the best decisions we could at the time.

One of the project's most significant strengths has been its ability to draw on deep expertise from the DHQ editorial team, which in turn derives partly from the fact that the focus of the project was on intelligent data modeling rather than on simple data capture. All of the editors approached the project as being in part an investigation of DHQ's citation universe, an unknown terrain to us and one in which we have an intense interest. The opportunity to inventory and model the range of cited materials—including everything from journal articles and book chapters to white papers, official reports, legal cases, private communications, tweets, blog posts, works of electronic literature, computer code, games, conference abstracts, works of fiction, manuscripts, newspaper articles, and dictionary entries—provided remarkable insight into the emergence of DH as a field and also into our own thinking about the mechanisms and purposes of scholarly citation. The editors also have a shared interest in data manipulation and data-driven work flows, so the practical challenges of the project (such as mechanisms for intelligent de-duplication) were framed as opportunities for the exercise of ingenuity. These motivations and interests continue to sustain this project after the conclusion of the grant funding. We also anticipate that the strong modeling of this data will make it more useful to third-party researchers.

## **Grant Products, Continuing Work, and Impact**

The most significant product arising from this grant is the bibliography itself, which is integrated into the DHQ interface but whose data can also be downloaded from the DHQ site. A secondary product is the set of supporting tools and systems (schemas, XSLT stylesheets, work flow) that enable DHQ to maintain and further develop this bibliography and its functions within the DHQ ecosystem. Another secondary product is the visualizations (and the analytic logic underlying them) that reveal patterns within the DHQ citations.

This project has a strong future trajectory for DHQ. One outcome of this project is a working system for bibliographic management in DHQ, and DHQ will now continue to use this system as part of our regular production work flow; hence we will naturally continue to expand the bibliography and groom it for quality. In addition, because DHQ is strongly committed to exploiting the journal's XML data and demonstrating the value of this data-driven approach to journal publishing, we will be seeking opportunities for further enhancements to both the data and the systems by which we expose it. As noted above, we plan a number of ongoing activities to bring this phase of development to completion. In addition, there are some longer-term projects that may arise from this work. In particular, we plan to solicit proposals for ways to exploit and analyze DHQ's data (including bibliographic data), possibly through microgrants in partnership with ADHO, and also through curricular opportunities such as the IVMOOC program mentioned above.

The long-term impact of this project on DHQ itself is likely to be very significant. As noted above, our previous system of bibliographic information was labor-intensive (since it required our encoding staff to copyedit and correct not only the content of each

citation, but also its punctuation and formatting which frequently diverged from DHQ's requested format) and duplicative (since many DHQ articles cite the same sources). Centralizing the bibliography not only does away with the most onerous parts of this work but also eliminates the duplication of information and the informational embarrassment of having the same work cited in different ways (since even conscientious authors may make different decisions concerning the inclusion of specific information, particularly in the case of less familiar genres such as white papers or conference proceedings). The satisfaction of maintaining a growing bibliography makes the labor of adding new entries much more tolerable. In addition, this data constitutes an important information resource that has great potential to enhance the DHQ interface. For example, we can enable readers of a given article to choose an item from its bibliography and discover all other DHQ articles that also cite the item, or to discover affinities between groups of DHQ articles based on their citation networks. Moreover, when we are able to expose this data to the public via an API, third party researchers may find additional ways to exploit the data (perhaps combining it or comparing it with other discipline-specific bibliographies). Through its impact on the DHQ interface and its potential to provide a valuable data resource to the public, this project raises DHQ's visibility in the digital humanities community and in related fields such as network science. Finally and perhaps most importantly, this project accomplished a task which can only be accomplished with funded labor, but which (once completed) lays the foundation for additional work that is interesting and lightweight enough to be done by volunteers or with small-scale funding such as microgrants. It thus served as a kind of gateway or enabling step which provides impetus for a much larger set of long-term effects

## Appendices

The appendices include the following items:

1. An XML code sample showing representative bibliographic entries encoded using the DHQ bibliographic markup.
2. A sample DHQ article encoded for publication, with a full bibliography showing the use of @key to point to the central bibliography (including handling of unlisted entries).
3. A screen shot of the side-by-side comparison view used to identify mismatched bibliographic entries during the deduplication and error correction phase of the project.
4. Internal documentation for the extraction and encoding of bibliographies from DHQ articles.
5. A final report by members of the IVMOOC working group describing their analysis of the DHQ bibliographic data.
6. The text and slides for a paper on DHQ (mentioning but not focused primarily on the bibliographic project) presented at DH2015 in Australia: "Challenges of an XML-based Open-Access Journal: Digital Humanities Quarterly," Julia Flanders, John Walsh, Wendell Piez, Melissa Terras. The text of this paper has been revised based on commentary and discussion in the conference session.

## ***Appendix 1: XML Code Sample***

This appendix contains an XML code sample showing representative bibliographic entries encoded using the DHQ bibliographic markup. The first set represent genres in common usage. The second set represent genres for which we are still considering the requirements and definitions.



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="../../../schema/dhqBiblio.rnc" type="application/relax-ng-compact-syntax"?>
<?xml-model href="../../../schema/dhqBiblio.sch" type="application/xml"
schematypens="http://purl.oclc.org/dsdl/schematron"?>
<?xml-styleSheet type="text/css" href="../../../css/dhqBiblio-formal.css" title="Formal" alternate="no"?>
```

```
<BiblioSet xmlns="http://digitalhumanities.org/dhq/ns/biblio"
xmlns:dhq="http://digitalhumanities.org/dhq/ns/biblio/util">
```

```
<Artwork ID="armstrong2005a" provenance="jm2014">
  <creator>
    <givenName>Kate</givenName>
    <familyName>Armstrong</familyName>
  </creator>
  <title>Grafik Dynamo</title>
  <date>2005</date>
  <url>http://www.turbulence.org/Works/dynamo/</url>
</Artwork>
```

```
<BlogEntry ID="manifesto2009" issuance="continuing" provenance="jm2014">
  <author>
    <corporateName>Humanities Blast</corporateName>
  </author>
  <title>Digital Humanities Manifesto 2.0</title>
  <date>2009</date>
  <url>http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf</url>
</BlogEntry>
```

```
<BookInSeries ID="morrison2000" issuance="monographic" provenance="jm2014">
  <author>
    <givenName>Grant</givenName>
    <familyName>Morrison</familyName>
  </author>
  <author>
    <givenName>J.G.</givenName>
    <familyName>Jones</familyName>
  </author>
  <title>Marvel Boy #4</title>
  <series issuance="continuing">
    <title>Marvel Boy</title>
    <publisher>Marvel Comics</publisher>
    <date>November 2000</date>
  </series>
</BookInSeries>
```

```
<BookSection ID="macdonald1998" provenance="jm2014" check-title-dupe="no">
  <author>
    <givenName>Sharon</givenName>
    <familyName>Macdonald</familyName>
  </author>
  <title>Introduction</title>
  <book issuance="monographic">
    <editor>
```

```
<givenName>Sharon</givenName>
<familyName>Macdonald</familyName>
</editor>
<title>Poetics of Display</title>
<place>London</place>
<publisher>Routledge</publisher>
<date>1998</date>
</book>
</BookSection>
```

```
<ConferencePaper ID="Marshall1998" issuance="monographic" provenance="jm2014">
  <author>
    <givenName>Catherine C.</givenName>
    <familyName>Marshall</familyName>
  </author>
  <title>Toward an ecology of hypertext annotation</title>
  <conference>
    <title>HyperText 98</title>
    <date>1998</date>
    <sponsor>ACM</sponsor>
  </conference>
  <startingPage>40</startingPage>
  <endingPage>49</endingPage>
</ConferencePaper>
```

```
<JournalArticle ID="macdonald2007" issuance="monographic" provenance="legacy">
  <author>
    <givenName>Moir</givenName>
    <familyName>MacDonald</familyName>
  </author>
  <title>Data Storage Policy Can't Be Enforced</title>
  <journal issuance="continuing">
    <title>University Affairs</title>
    <volume>4</volume>
    <date>June 2007</date>
  </journal>
  <url>http://www.universityaffairs.ca/data-storage-policy-cant-be-enforced.aspx</url>
</JournalArticle>
```

```
<Thesis ID="martiradonna2004" issuance="monographic" provenance="jm2014">
  <author>
    <givenName>V.</givenName>
    <familyName>Martiradonna</familyName>
  </author>
  <title>La codifica elettronica dei testi. Un caso di studio</title>
  <place>Tesi di laurea in Lettere, Facoltà di Scienze Umanistiche, Università di Roma La Sapienza</place>
  <date>2003-2004</date>
  <note>Relatore: D. Fiormonte.</note>
</Thesis>
```

```
<VideoGame ID="montfort2000" issuance="monographic" provenance="jm2014">
  <creator>
```

<givenName>Nick</givenName>  
<familyName>Montfort</familyName>  
</creator>  
<title>Ad Verbum</title>  
<date>2000</date>  
<url><http://www.wurb.com/if/game/912></url>  
</VideoGame>

<WebSite ID="maher2005" issuance="continuing">  
<author>  
<givenName>Jimmy</givenName>  
<familyName>Maher</familyName>  
</author>  
<title>Review of Rune Berg's The Isle of the Cult</title>  
<date>2005</date>  
<url><http://www.sparkynet.com/spag/i.html#isle></url>  
</WebSite>

<!-- further proposed genres pending review -->

<ArchivalItem ID="castle" issuance="monographic" provenance="jm2014">  
<title>The Castle of Perseverance Map</title>  
<additionalTitle>The Castle of Perserverance</additionalTitle>  
<date/>  
<note>MS. Folger Shakespeare Library, Washington. Shelfmark V.a.354. 191v. Image ID 1207-42.</note>  
</ArchivalItem>

<CourtCase ID="baker1879" issuance="monographic" provenance="jm2014">  
<title>Baker v. Selden</title>  
<date>1879</date>  
<note>101 U.S. 99</note>  
</CourtCase>

<Legal ID="constitution" issuance="continuing" provenance="jm2014">  
<title>U.S. Constitution, Article 1, Section 8</title>  
</Legal>

<Proposal ID="IATHa" issuance="monographic">  
<author>  
<corporateName>Institute for Advanced Technologies in the Humanities (IATH)</corporateName>  
</author>  
<title>NEH Proposal</title>  
<additionalTitle> SNAC: The Social Networks and Archival Context Project. </additionalTitle>  
<url>[http://socialarchive.iath.virginia.edu/NEH\\_proposal\\_narrative.pdf](http://socialarchive.iath.virginia.edu/NEH_proposal_narrative.pdf)</url>  
<note>Accessed April 15, 2012</note>  
</Proposal>

<Report ID="terras2006a" issuance="monographic" provenance="jm2014">  
<author>  
<givenName>Melissa</givenName>  
<familyName>Terras</familyName>

</author>

<title>he Researching e-Science Analysis of Census Holdings Project: Final Report to  
AHRC</title>

<date>2006</date>

<url>www.ucl.ac.uk/reach/</url>

<note>AHRC e-Science Workshop scheme</note>

</Report>

</BiblioSet>

## ***Appendix 2: Sample DHQ Article***

This appendix contains a sample DHQ article encoded for publication, with a full bibliography showing the use of @key to point to the central bibliography (including handling of unlisted entries using @key="[unlisted]").

```

<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="../../../common/schema/DHQauthor-TEI.rng" type="xml"?>
<?oxygen SCHSchema="../../../common/schema/dhqTEI-ready.sch"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:dhq="http://www.digitalhumanities.org/ns/dhq"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:cc="http://web.resource.org/cc/">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>The Technical Evolution of Vannevar Bush's Memex </title>
        <author>Belinda Barnet</author>
        <dhq:authorInfo>
          <dhq:author_name>Belinda <dhq:family>Barnet</dhq:family>
          </dhq:author_name>
          <dhq:affiliation>Swinburne University of Technology, Melbourne</dhq:affiliation>
          <email>belinda.barnet at gmail.com</email>
          <dhq:bio>
            <p>Belinda Barnet is Lecturer in Media and Communications at Swinburne University,
              Melbourne. Prior to her appointment at Swinburne she worked at Ericsson
              Australia, where she managed the development of 3G mobile content services and
              developed an obsession with technical evolution. Belinda did her PhD on the
              history of hypertext at the University of New South Wales, and has research
              interests in digital media, digital art, convergent journalism and the mobile
              internet. She has published widely on new media theory and culture.</p>
            </dhq:bio>
          </dhq:authorInfo>
        </titleStmt>
        <publicationStmt>
          <idno type="DHQarticle-id">000015</idno>
          <idno type="volume">002</idno>
          <idno type="issue">1</idno>
          <dhq:articleType>article</dhq:articleType>
          <date when="2008-06-21">21 June 2008</date>
          <availability>
            <cc:License xmlns="http://digitalhumanities.org/DHQ/namespace"
              rdf:about="http://creativecommons.org/licenses/by-nc-nd/2.5/" />
            </availability>
          </publicationStmt>
          <sourceDesc>
            <p>Authored for DHQ; migrated from original DHQauthor format</p>
          </sourceDesc>
        </fileDesc>
        <encodingDesc>
          <classDecl>
            <taxonomy xml:id="dhq_keywords">
              <bibl>DHQ classification scheme; full list available in the <ref
                target="http://www.digitalhumanities.org/dhq/taxonomy.xml">DHQ keyword
                taxonomy</ref>
              </bibl>
            </taxonomy>
            <taxonomy xml:id="authorial_keywords">
              <bibl>Keywords supplied by author; no controlled vocabulary</bibl>
            </taxonomy>
          </classDecl>

```

```

</encodingDesc>
<profileDesc>
  <langUsage>
    <language ident="en"/>
  </langUsage>
</profileDesc>
<revisionDesc>
  <change when="2008-06-14" who="jif">Added final metadata, bio and abstract, publication
    statement, proofreading corrections. Restored # to targets where it was missing, for
    consistency.</change>
  <change when="2008-04-26" who="Flanders">Encoded document</change>
  <change when="2008-05-24" who="Ashwini"> Added date, id, issue, vol attributes to root
    element, revised encoding of the change element dated 2008-04-26, removed "#" from
    target attribute of ref element, encoded external links as xref in the listBibl, removed
    top xsl declaration </change>
  <change when="2008-06-16" who="Ashwini">Updated revisionDesc format, added details to
    publicationStmt, changed xref to ref for validation with new schema, added some missing
    "#" to target attribute and removed "##".</change>
  <change when="2008-06-20" who="jif">Changed email address, made authorial changes.</change>
</revisionDesc>
</teiHeader>
<text>
  <front>
    <dhq:abstract>
      <p>This article describes the evolution of the design of Vannevar Bush's Memex, tracing
        its roots in Bush's earlier work with analog computing machines, and his
        understanding of the technique of associative memory. It argues that Memex was the
        product of a particular engineering culture, and that the machines that preceded
        Memex — the Differential Analyzer and the Selector in particular — helped engender
        this culture, and the discourse of analogue computing itself.</p>
    </dhq:abstract>
    <dhq:teaser>
      <p>Can we say that technical machines have their own genealogies, their own evolutionary
        dynamic?</p>
    </dhq:teaser>
  </front>
  <body>
    <div>
      <head>Introduction: Technical Evolution</head>
      <cit>
        <quote rend="block">The key difference [between material cultural evolution and
          biological evolution] is that biological systems predominantly have
          <q>vertical</q> transmission of genetically ensconced information, meaning
          parents to offspring... Not so in material cultural systems, where horizontal
          transfer is rife — and arguably the more important dynamic .</quote>
        <ref target="#eldredge2004">Paleontologist Dr. Niles Eldredge, interview with the
          author</ref>
      </cit>
      <p>Since the early days of Darwinism, analogies have been drawn between biological
        evolution and the evolution of technical objects and systems. It is obvious that
        technologies change over time; we can see this in the fact that technologies come in
        generations; they adapt and adopt characteristics over time, <cit>
          <quote rend="inline">one suppressing the other as it becomes obsolete</quote>

```

[guattari1995](#) loc="40"/>

</cit>. The technical artefact constitutes a series of objects, a <q>lineage</q> or a line. From the middle of the nineteenth century on, writers have been remarking on this basic analogy – and on the alarming rate at which technological change is accelerating. But as Eldredge points out, the analogy can only go so far; technological systems are not like biological systems in a number of important ways, most obviously the fact that they are the products of conscious design. Unlike biological organisms, technical objects are **invented**. </p>

<p>Inventors learn by experience and experiment, and they learn by watching other machines work in the form of technical prototypes. They also copy and <q>transfer</q> ideas and techniques between machines, co-opting innovations at a whim. Technological innovation thus has Lamarckian features, which are forbidden in biology <ptr target="#ziman2003" loc="5"/>. Inventors can borrow ideas from contemporary technologies, or even from the past. There is no <q>extinction</q> in technological evolution: ideas, designs and innovations can be co-opted and transferred both retroactively and laterally. This retroactive and lateral <q>transfer</q> of innovations is what distinguishes technical evolution from biological evolution, which is characterised by vertical transfer (parents to offspring). As the American paleontologist Niles Eldredge observed in an interview with the author,</p>

<cit>

<quote rend="block">Makers copy each other, patents affording only fleeting protection. Thus, instead of the neatly bifurcating trees [you see in biological evolution], you find what is best described as "networks"-consisting of an historical signal of what came before what, obscured often to the point of undetectability by this lateral transfer of subsequent ideas .</quote>

<ref target="#eldredge2004">Niles Eldredge, interview with the author</ref>

</cit>

<p>Can we say that technical machines have their own genealogies, their own evolutionary dynamic? It is my contention that we can, and I have argued elsewhere that in order to tell the story of a machine, one must trace the path of these transferrals, paying particular attention to technical **prototypes** and to also to

**techniques**, or ways of doing things. A good working prototype can send shockwaves throughout an engineering community, and often inspires a host of new machines in quick succession. Similarly, an effective technique (for example, storing and retrieving information associatively) can spread between innovations rapidly.</p>

<p>In this article I will be telling the story of particular technical machine –

Vannevar Bush's Memex. Memex was an electro-mechanical device designed in the 1930's to provide easy access to information stored associatively on microfilm. It is often hailed as the precursor to hypertext and the web. Linda C. Smith undertook a comprehensive citation context analysis of literary and scientific articles produced after the 1945 publication of Bush's article on the device, <title rend="quotes">As

We May Think</title> in the <title rend="italic">Atlantic Monthly</title>. She found that there is a conviction, without dissent, that modern hypertext is traceable to this article <ptr target="#smith1991" loc="265"/>. In each decade since the Memex design was published, commentators have not only lauded it as vision, but also asserted that <cit>

<quote rend="inline">technology [has] finally caught up with this vision</quote>

<ptr target="#smith1991" loc="278"/>

</cit>. For all the excitement, it is important to remember that Memex was never actually built; it exists entirely on paper. Because the design was first published in the summer of 1945, at the end of a war effort and with the birth of computers, theorists have often associated it with the post-War information boom. In fact, Bush had been writing about it since the early 1930s, and the Memex paper went through



several different versions. </p>

<p>The social and cultural influence of Bush's inventions are well known, and his political role in the development of the atomic bomb are also well known. What is not so well known is the way the Memex came about as a result of both Bush's earlier work with analog computing machines, and his understanding of the <quote rend="inline">mechanism</quote> or technique of associative memory. I would like to show that Memex was the product of a particular engineering culture, and that the machines that preceded Memex — the Differential Analyzer and the Selector in particular — helped engender this culture, and the discourse of analogue computing, in the first place. The artefacts of engineering, particularly in the context of a school such as MIT, are themselves productive of new techniques and new engineering paradigms. Prototype technologies create cultures of use around themselves; they create new techniques and new methods that were unthinkable prior to the technology. This was especially so for the Analyzer. </p>

<cit>

<quote rend="block">In the context of the early 20th-century engineering school, the analyzers were not only tools but paradigms, and they taught mathematics and method and modeled the character of engineering.</quote>

<ptr target="#owens1991" loc="6"/>

</cit>

<p>Bush transferred technologies directly from the Analyzer and also the Selector into the design of Memex. I will trace this transfer in the first section. He also transferred an electro-mechanical model of human associative memory from the nascent science of cybernetics, which he was exposed to at MIT, into Memex. We will explore this in the second section. In both cases, we will be paying particular attention to the structure and architecture of the technologies concerned.</p>

<p>The idea that technical artefacts evolve in this way, by the transfer of both technical innovations (for example, microfilm) and techniques (for example, association as a storage technique), was popularised by French technology historian Bertrand Gille. I will be mobilising Gille's theories here as I trace the evolution of the Memex design. We will begin with Bush's first analogue computer, the Differential Analyzer.</p>

</div>

<div>

<head>The Analyzer and the Selector</head>

<p>The Differential Analyzer was a giant, electromechanical gear and shaft machine which was put to work during the war calculating artillery ranging tables and the profiles of radar antennas. In the late 1930s and early 1940s, it was <cit>

<quote rend="inline">the most important computer in existence in the US</quote>

<ptr target="#owens1991" loc="3"/>

</cit>. Before this time, the word <q>computer</q> had meant a large group of mostly female humans performing equations by hand or on limited mechanical calculators. The Analyzer evaluated and solved these equations by mechanical integration. It created a small revolution at MIT. Many of the people who worked on the machine (e.g. Harold Hazen, Gordon Brown, Claude Shannon) later made contributions to feedback control, information theory, and computing <ptr target="#mindell2000"/>. The machine was a huge success which brought prestige and a flood of federal money to MIT and Bush. </p>

<p>However, by the spring of 1950, the Analyzer was gathering dust in a storeroom — the project had died. Why did it fail? Why did the world's most important analogue computer end up in a back room within five years? This story will itself be related to why Memex was never built; research into analogue computing technology in the interwar years, the Analyzer in particular, contributed to the rise of digital computing. It demonstrated that <emph>machines could automate the calculus</emph>,</p>

that machines could automate human cognitive techniques.

The decade between the Great War and the Depression was a bull market for engineering [Enrolment in the MIT Electrical Engineering Department almost doubled in this period, and the decade witnessed the rapid expansion of graduate programs. The interwar years found corporate and philanthropic donors more willing to fund research and development within engineering departments, and there were serious problems to be worked on generated by communications failures during the Great War. In particular, engineers were trying to predict the operating characteristics of power-transmission lines, long-distance telephone lines, commercial radio and other communications technologies \(Beniger calls this the early period of the Control Revolution](#) [MIT's Engineering Department undertook a major assault on the mathematical study of long-distance lines.](#)

Of particular interest to the engineers was the Carson equation for transmission lines. This was a simple equation, but it required intensive mathematical integration to solve.

Early in 1925 Bush suggested to his Graduate Student Herbert Stewart that he devise a machine to facilitate the recording of the areas needed for the Carson equation ... [and a colleague] suggested that Stewart interpret the equation electrically rather than mechanically.

So the equation was transferred to an electro-mechanical device: the Product Intergraph. Many of the early analogue computers that followed Bush's machines were designed to automate existing mathematical equations. This particular machine physically mirrored the equation itself. It incorporated the use of a mechanical integrator to record the areas under the curves (and thus the integrals), which was

... in essence a variable-speed gear, and took the form of a rotating horizontal disk on which a small knife-edged wheel rested. The wheel was driven by friction, and the gear ratio was altered by varying the distance of the wheel from the axis of rotation of the disk.

A second version of this machine incorporated two wheel-and-disc integrators, and it was a great success. Bush observed the success of the machine, and particularly the later incorporation of the two wheel-and-disc integrators, and decided to make a larger one, with more integrators and a more general application than the Carson equation. By the fall of 1928, Bush had secured funds from MIT to build a new machine. He called it the Differential Analyzer, after an earlier device proposed by Lord Kelvin which might externalise the calculus and

mechanically integrate its solution

As Bertrand Gille observes, a large part of technical invention occurs by transfer, whereby the functioning of a structure is analogically transposed onto another structure, or the same structure is generalised outwards

This is what happened with the Analyzer — Bush saw the outline of such a machine in the Product Intergraph. The Differential Analyzer was rapidly assembled in 1930, and part of the reason it was so quickly done was that it incorporated a number of existing engineering developments, particularly a device called a torque amplifier, designed by Niemann. But the disk integrator, a technology borrowed from the Product Intergraph, was the heart of the

Analyzer and the means by which it performed its calculations. When combined with the torque amplifier, the Analyzer was [cited](#)

[essentially an elegant, dynamical, mechanical model of the differential equation](#)

[#owens1991" loc="14"/>](#)

[. Although Lord Kelvin had suggested such a machine previously, Bush was the first to build it on such a large scale, and it happened at a time when there was a general and urgent need for such precision. It created a small revolution at MIT.](#)

In engineering science, there is an emphasis on working prototypes or

[deliverables](#). As Professor of Computer Science Andries van Dam put it in an interview with the author, when engineers talk about work, they mean [cited](#)

[work in the sense of machines, software, algorithms, things that are concrete](#)

[quote](#)

[#vandam1999"/>](#)

[. This emphasis on concrete work was the same in Bush's time. Bush had delivered something which had been previously only been dreamed about; this meant that others could come to the laboratory and learn by observing the machine, by watching it integrate, by imagining other applications. A working prototype is different to a dream or white paper — it actually creates its own milieu, it](#)

[teaches those who use it](#) about the possibilities it contains and its material technical limits. Bush himself recognised this, and believed that those who used the machine acquired what he called a [mechanical calculus](#), an internalised knowledge of the machine. When the army wanted to build their own machine at the Aberdeen Proving Ground, he sent them a mechanic who had helped construct the Analyzer. The army wanted to pay the man machinist's wages; Bush insisted he be hired as a consultant [#owens1991" loc="24"/>](#). [cited](#)

[I never consciously taught this man any part of the subject of differential equations; but in building that machine, managing it, he learned what differential equations were himself ... \[it\] was interesting to discuss the subject with him because he had learned the calculus in mechanical terms — a strange approach, and yet he understood it. That is, he did not understand it in any formal sense, he understood the fundamentals; he had it under his skin.](#)

[#bush1970">Bush 1970, 262 cited in Owens 1991, 24](#)

[cited](#)

Watching the Analyzer work did more than just teach people about the calculus. It also taught people about what might be possible for mechanical calculation — for

[analogue computers](#). Several laboratories asked for plans, and duplicates were set up at the US Army's Ballistic Research Laboratory, in Maryland, and at the Moore School of Electrical Engineering at the University of Pennsylvania

[#shurkin1996" loc="99"/>](#). The machine assembled at the Moore school was much larger than the MIT machine, and the engineers had the advantage of being able to learn from the mistakes and limits of the MIT machine [loc="102"](#)

[#shurkin1996"/>](#). Bush also created several more Analyzers, and in 1936 the Rockefeller Foundation awarded MIT \$85,000 to build the Rockefeller Differential Analyzer [#owens1991" loc="17"/>](#). This provided more opportunities for graduate research, and brought prestige and a flood of funding to MIT. [p>](#)

[cited](#)

[But what is interesting about the Rockefeller Differential Analyzer is what remained the same. Electrically or not, automatically or not, the newest edition of Bush's analyzer still interpreted mathematics in terms of](#)

mechanical rotations, still depended on expertly machined wheel-and-disc integrators, and still drew its answers as curves.</quote>

<ptr loc="32" target="#owens1991"/>

</cit>

<p>Its technical processes remained the same. It was an analogue device, and it literally turned around a central analogy: the rotation of the wheel shall be the area under the graph (and thus the integrals). The Analyzer directly mirrored the task at hand; there was a mathematical transparency to it which at once held observers captive and promoted, in its very workings, the <cit>

<quote rend="inline">language of early 20th-century engineering</quote>

<ptr target="#owens1991" loc="32"/>

</cit>. There were visitors to the lab, and military and corporate representatives that would watch the machine turn its motions. It seemed the adumbration of future technology. Harold Hazen, the head of the Electrical Engineering Department in 1940 predicted the Analyzer would <cit>

<quote rend="inline">mark the start of a new era in mechanized calculus</quote>

<ref target="#hazen1940">Hazen 1940, 101 cited in Owens 1991, 4</ref>

</cit>. Analogue technology held much promise, especially for military computation — and the Analyzer had created a new era. The entire direction and culture of the MIT lab changed around this machine to woo sponsors <ptr target="#nyce1991" loc="39"/>. In the late 1930s the department became the Center of Analysis for Calculating Machines.</p>

<p>Many of the Analyzers built in the 1930s were built using military funds. The creation of the first Analyzer, and Bush's <emph>promotion</emph> of it as a calculation device for ballistic analysis, had created a link between the military and engineering science at MIT which was to endure for over thirty years. Manuel De Landa (1994) puts great emphasis in his work on this connection, particularly as it was further developed during WWII. As he puts it, Bush created a <quote rend="inline">

>bridge</quote> between the engineers and the military, he <cit>

<quote rend="inline">connected scientists to the blueprints of generals and admirals</quote>

<ptr loc="119" target="#delanda1994"/>

</cit>, and this relationship would grow infinitely stronger during WWII.

Institutions that had previously occupied exclusive ground such as physics and military intelligence had begun communicating in the late 1930s, <cit>

<quote rend="inline">communities often suspicious of one another: the inventors and the scientists on the one side and the warriors on the other</quote>

<ptr target="#delanda1994" loc="36"/>

</cit>. </p>

<p>This paper has been arguing that the Analyzer <foreign>qua</foreign> technical artefact accomplished something equally important: as a prototype, it demonstrated the potential of analogue computing technology for analysis, and engendered an engineering culture around itself that took the machine to be a teacher. This is why, even after the obsolescence of the Analyzer, it was kept around at MIT for its educational value <ptr target="#owens1991" loc="23"/>. It demonstrated that machines could automate the calculus, and that machines could mirror human tasks in an elegant fashion: something which required proof in steel and brass. The <q>aura</q> generated by the Analyzer as prototype was not lost on the military.</p>

<p>In 1935, the Navy came to Bush for advice on machines to crack coding devices like the new Japanese cipher machines <ptr target="#burke1991" loc="147"/>. They wanted a long-term project that would give the United States the most technically advanced cryptanalytic capabilities in the world, a super-fast machine to count the coincidences of letters in two messages or copies of a single message. Bush assembled

a research team for this project that included Claude Shannon, one of the early information theorists and a significant part of the emerging cybernetics community

[nyce1991](#) loc="40"/>.

There were three new technologies emerging at the time which handled information: photoelectricity, microfilm and digital electronics.

All three were just emerging, but, unlike the fragile magnetic recording his students were exploring, they appeared to be ready to use in calculation machines. Microfilm would provide ultra-fast input and inexpensive mass-memory, photoelectricity would allow high-speed sensing and reproduction, and digital electronics would allow astonishingly fast and inexpensive control and calculation.

[burke1991](#) loc="147"/>

Bush transferred these three technologies to the new design. This decision was not pure genius on his part; they were perfect analogues for a popular conception of how the brain worked at the time. The scientific community at MIT were developing a pronounced interest in man-machine analogues, and although Claude Shannon had not yet published his information theory it was already being formulated, and there was much discussion around MIT about how the brain might process information in the manner of an analogue machine. Bush thought and designed in terms of analogies between brain and machine, electricity and information. This was also the central research agenda of Norbert Weiner and Warren McCulloch, both at MIT, who were at the time [working on parallels they saw between neural structure and process and computation](#) ([nyce1991](#) loc="63"/>; see also [hayles1999](#)). To Bush and Shannon, microfilm and photoelectricity seemed perfect analogues to the [electrical relay circuits and neural substrates of the human brain](#) and their capacities for managing information.

Bush called this machine the Comparator — it was to do the hard work of comparing text and letters for the humble human mind. Like the analytic machines before it and all other technical machines being built at the time, this was an analogue device; it directly mirrored the task at hand on a mechanical level. In this case, it directly mirrored the operations of [searching](#) and [associating](#) on a mechanical level, and, Bush believed, it mirrored the operations of the human mind and memory. Bush began the project in mid-1937, while he was working on the Rockefeller Analyzer, and agreed to deliver a code-cracking device based on these technologies by the next summer [burke1991](#) loc="147"/>.

But immediately, there were problems in its development. Technical objects often depart from their fabricating intention; sometimes because they are [used](#) differently to what they were invented for, and sometimes because the

[technology itself](#) breaks down. Microfilm did not behave the way Bush wanted it to. As a material it was very fragile, sensitive to light and heat, and tore easily; it had too many [bugs](#). It was decided to use paper tape with minute holes, although paper was only one-twentieth as effective as microfilm [burke1991](#) loc="147"/>. There were subsequent problems with this technology — paper itself is flimsy, and it refused to work well for long periods intact. There were also problems shifting the optical reader between the two message tapes. Bush was working on the Analyzer at the time, and didn't have the resources to fix these components effectively. By the time the Comparator was turned over to the Navy, it was very unreliable, and didn't even start up when it was unpacked in Washington [burke1991](#) loc="148"/>. The Comparator prototype ended up gathering dust in a Navy storeroom, but much of the architecture was transferred to subsequent designs.



<p>By this time, Bush had also started work on the Memex design. He transferred much of the architecture from the Comparator, including photoelectrical components, an optical reader and microfilm. In tune with the times, Bush had developed a fascination for microfilm in particular as an information storage technology, and although it had failed to work properly in the Comparator, he wanted to try it again. It would appear as the central technology in the Rapid Selector and also in the Memex design.</p>

<p>In the 1930s, many believed that microfilm would make information universally accessible and thus spark an intellectual revolution (<ptr target="#farkas-conn1990" loc="16–22"/>, cited in <ptr target="#nyce1991" loc="49"/>). Like many others, he had been enthusiastically exploring its potential in his writing <ptr target="#bush1933"/>, <ptr target="#bush1939"/> as well as the Comparator; the Encyclopaedia Britannica <quote rend="inline">could be reduced to the volume of a matchbox. A library of a million volumes could be compressed into one end of a desk</quote> he wrote <ptr target="#bush1945" loc="93"/>. In 1938, H.G. Wells even wrote about a <quote rend="inline">Permanent World Encyclopaedia</quote> or Planetary Memory that would carry all the world's knowledge. It was based on microfilm. </p>

<cit>

<quote rend="block">By means of microfilm, the rarest and most intricate documents and articles can be studied now at first hand, simultaneously in a score of projection rooms. There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas, achievements, to the creation, that is, of a complete planetary memory for all mankind.</quote>

<dhq:citRef>

<ptr target="#wells1938"/>, cited in <ptr target="#nyce1991" loc="50"/>

</dhq:citRef>

</cit>

<p>Microfilm promised <emph>faithful reproduction</emph> as well as miniaturisation. It was state-of-the-art technology, and not only did it seem the perfect analogy for material stored in the neural substrate of the human brain, it seemed to have a certain permanence the brain lacked. Bush put together a proposal for a new microfilm selection device, based on the architecture of the Comparator, in 1937. Its stated research agenda and intention was</p>

<cit>

<quote rend="block">

<list>

<item>Construction of experimental equipment to test the feasibility of a device which would search reels of coded microfilm at high speed and which would copy selected frames on the fly, for printout and use.</item>

<item>Investigation of the practical utility of such equipment by experimental use in a library.</item>

<item>Further development aimed at exploration of the possibilities for introducing such equipment into libraries generally.</item>

</list>

</quote>

<ref target="#bagg1961">Bagg and Stevens 1961, cited in Nyce 1991, 41</ref>

</cit>

<p>Corporate funding was secured for the Selector by pitching it as a microfilm machine to modernise the library <ptr target="#nyce1991" loc="41"/>. Abstracts of documents were to be captured by this new technology and reduced in size by a factor of 25. As with the Comparator, long rolls of this film were to be spun past a photoelectric sensing station. If a match occurred between the code submitted by a researcher and

the abstract codes attached to this film [<ptr target="#burke1991" loc="151"/>](#), the researcher was presented with the article itself and any articles previously associated with it. This was to be used in a public library, and unlike his nascent idea concerning Memex, he wanted to tailor it to commercial and government record-keeping markets. [</p>](#)

[<p>](#)Bush considered the Selector as a step towards the mechanised control of scientific information, which was of immediate concern to him as a scientist. According to him, the fate of the nation depended on the effective management of these ideas lest they be lost in a brewing data storm. Progress in information management was not only inevitable, it was [<cit>](#)

[<quote rend="inline">](#)essential if the nation is to be strong[</quote>](#)

[<ptr target="#bush1970" loc="149"/>](#)

[</cit>](#). This was his fabricating intention. He had been looking for support for a Memex-like device for years, but after the failure of the Comparator, finding funds for this [<quote rend="inline">](#)library of the future[</quote>](#) was very hard [<ptr target="#burke1991" loc="149"/>](#). Then in 1938, Bush received funding from the National Cash Register Company and the Eastman Kodak Company for the development of an apparatus for rapid selection, and he began to transfer the architecture from the Comparator across to the new design. [</p>](#)

[<p>](#)But as Burke writes, the technology of microfilm and the tape-scanners began to impose their technical limitations; [<cit>](#)

[<quote rend="block">](#)[a]lmost as soon as it was begun, the Selector project drifted away from its original purpose and began to show some telling weaknesses ... Bush planned to spin long rolls of 35mm film containing the codes and abstracts past a photoelectric sensing station so fast, at speeds of six feet per second, that 60,000 items could be tested in one minute. This was at least one hundred-fifty times faster than the mechanical tabulator.[</quote>](#)

[<ptr target="#burke1991" loc="150"/>](#)

[</cit>](#)

[</p>](#)

[<p>](#)The Selector's scanning station was similar to that used in the Comparator. But in the Selector, the card containing the code of interest to the researcher would be stationary. Bush and others associated with the project [<cit>](#)

[<quote rend="inline">](#)were so entranced with the speed of microfilm tape that little attention was paid to coding schemes[</quote>](#)

[<ptr target="#burke1991" loc="151"/>](#)

[</cit>](#), and when Bush handed the project over to three of his researchers, John Howard, Lawrence Steinhardt and John Coombs, it was floundering. After three more years of intensive research and experimentation with microfilm, Howard had to inform the Navy that the machine would not work [<ptr target="#burke1991" loc="149"/>](#). Microfilm, claimed Howard, would deform at such speeds and could not be aligned so that coincidences could be identified. Microfilm warps under heat, and it cannot take great strain or tension without distorting. [</p>](#)

[<p>](#)Solutions were suggested (among them slowing down the machine, and checking abstracts before they were used) [<ptr target="#burke1991" loc="154"/>](#), but none of these were particularly effective, and a working machine wasn't ready until the fall of 1943. At one stage, because of an emergency problem with Japanese codes, it was rushed to Washington — but because it was so unreliable, it went straight back into storage. So many parts were pulled out that the machine was never again operable [<ptr target="#burke1991" loc="158"/>](#). In 1998, the Selector made Bruce Sterling's Dead Media List, consigned forever to a lineage of failed technologies. Microfilm did not behave the way Bush and his team wanted it to. It had its own material limits, and these didn't support speed of access. [</p>](#)

<p>In the evolution of any machine, there will be internal limits generated by the behaviour of the technology itself; Gille calls these <quote rend="inline">endogenous</quote> limits <ptr target="#gille1986"/>. Endogenous limits are encountered only in practice — they effect the actual <emph>implementation</emph> of an idea. In engineering practice, these failures can teach inventors about the material potentials of the technology as well. The Memex design altered significantly through the 1950s; Bush had learned from the technical failures he was encountering. But most noticeable of all, Bush stopped talking about microfilm and about hardware.</p>

<cit>

<quote rend="block">By the 1960's the project and machine failures associated with the Selector, it seems, made it difficult for Bush to think about Memex in concrete terms.</quote>

<ptr target="#burke1991" loc="161"/>

</cit>

<p>The Analyzer, meanwhile, was being used extensively during WWII for ballistic analysis and calculation. Wartime security prevented its public announcement until 1945, when it was hailed by the press as a great <quote rend="inline">electromechanical brain</quote> ready to advance science by freeing it from the pick-and-shovel work of mathematics (<ref target="#owens1991"><title rend="italic">Life</title> magazine, cited by Owens 1991, 3</ref>). It had created an entire culture around itself. But by the mid-1940s, the enthusiasm had died down; the machine seemed to pale beside the new generation of digital machines. The war had also released an unprecedented sum of money into MIT and spawned numerous other new laboratories. It <cit>

<quote rend="inline">ushered in a variety of new computation tasks, in the field of large-volume data analysis and real-time operation, which were beyond the capacity of the Rockefeller instrument</quote>

<ptr target="#owens1991" loc="5"/>

</cit>. By 1950, the Analyzer had become an antique, conferred to back-room storage. </p>

<p>What happened? The reasons The Analyzer fell into disuse were quite different to the Selector; its limits were <emph>exogenous</emph> to the technical machine itself. They were related to a fundamental paradigm shift within computing, from analogue to digital. According to Gille, the birth of a new technical system is rapid and unforeseeable; new technical systems are born with the limits of the old technical systems, and the period of change is brutal, fast and discontinuous. In 1950, Warren Weaver and Samuel Caldwell met to discuss the Analyzer and the analogue computing program it had inspired at MIT, a large program which had become out of date more swiftly than anyone could have imagined. They noted that in 1936, no one could have expected that within ten years the whole field of <q>computer science</q> would so quickly overtake Bush's project (<ref target="#weaver1950a">Weaver and Caldwell</ref>, cited in <ptr target="#owens1991" loc="4"/>). Bush, and the department at MIT which had formed itself around the Analyzer and analogue computing, had been left behind.</p>

<p>I do not have the space here to trace the evolution of digital computing at this time in the US and the UK — excellent accounts have already been written by <ptr target="#beniger1986"/>, <ptr target="#shurkin1996"/>, <ptr target="#ceruzzi1998"/>, <ptr target="#edwards1997"/> and <ptr target="#delanda1994"/> to name a few. All we need to realise at this point is that the period between 1945 and 1967, the years between the publication of the first and the final versions of the Memex essays respectively, had witnessed enormous change. The period saw not only the rise of digital computing, beginning with the construction of a few machines in the post-war period and developing into widespread mainframe processing for American business, it



also saw the explosive growth of commercial television [spar2001](#)<sup>194</sup>, and the beginnings of satellite broadcasting [spar2001](#)<sup>197</sup>. As Beniger sees it, the world had discovered information as a means of control [beniger1986](#)<sup>vii</sup>.

It is important to understand, however, that Bush was not a part of this revolution. He had not been trained in digital computation or information theory, and knew little about the emerging field of digital computing. He was immersed in a different technical system: analogue machines interpreted mathematics in terms of mechanical rotations, storage and memory as a physical [holding](#) of information, and drew their answers as curves. They directly mirrored the operations of the calculus. Warren Weaver expressed his regret over the passing of analogue machines and the Analyzer in a letter to the director of MIT's Center of Analysis: [cit](#)

[rend="block"](#)It seems rather a pity not to have around such a place as MIT a really impressive Analogue computer; for there is a vividness and directness of meaning of the electrical and mechanical processes involved ... which can hardly fail, I would think, to have a very considerable educational value.[/quote](#)

[ref target="#owens1991"](#)Weaver, cited in Owens 1991, 5[/ref](#)

[/cit](#)

[/p](#)

The passing away of analogue computing was the passing away of an ethos: machines as mirrors of mathematical tasks. But Bush and Memex remained in the analogue era; in all versions of the Memex essay, his goal remained the same: [cit](#)

[quote rend="inline"](#)he sought to develop a machine that mirrored and recorded the patterns of the human brain[/quote](#)

[ptr target="#nyce1991" loc="123"/>](#)

[/cit](#), even when this era of direct reflection and analogy in mechanical workings had passed. [/p](#)

Technological evolution moves faster than our ability to adjust to its changes. More precisely, it moves faster than the [emph](#)techniques[/emph](#) that it engenders and the culture it forms around itself. Bush expressed some regret over this speed of passage near the end of his life, or, perhaps, sadness over the obsolescence of his own engineering techniques. [/p](#)

[cit](#)

[quote rend="block"](#)The trend had turned in the direction of digital machines, a whole new generation had taken hold. If I mixed with it, I could not possibly catch up with new techniques, and I did not intend to look foolish.[/quote](#)

[ptr target="#bush1970" loc="208"/>](#)

[/cit](#)

[/div](#)

[div](#)

[head](#)Human Associative Memory and Biological-Mechanical Analogues[/head](#)

[cit](#)

[quote rend="block"](#)There is another revolution under way, and it is far more important and significant than [the industrial revolution]. It might be called the mental revolution.[/quote](#)

[ptr target="#bush1959" loc="165"/>](#)

[/cit](#)

We now turn to Bush's fascination with, and exposure to, new models of human associative memory gaining current in his time. Bush thought and designed his machines in terms of biological-mechanical analogues; he sought a symbiosis between

[q](#)natural[/q](#) human thought and his thinking machines. [/p](#)

As Nyce and Kahn observe, in all versions of the Memex essay (1939, 1945, 1967), Bush

begins his thesis by explaining the dire problem we face in confronting the great mass of the human record, criticising the way information was then organised [target="#nyce1991" loc="56"/>. He then goes on to explain the reason why this form of organisation doesn't work: it is \*\*artificial\*\*. Information should be organised by association — this is how the mind works. If we fashion our information systems after this mechanism, they will be truly revolutionary.](#)

<cit>

<quote rend="block">

<p>Our ineptitude at getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can only be found in one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.</p>

<p>The human mind does not work that way. It operates by association. With one item in grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain.</p>

</quote>

<ref target="#bush1939">Bush 1939, 1945, 1967</ref>

</cit>

<p>These paragraphs were important enough that they appeared verbatim in all versions of the Memex essay — 1939, 1945 and 1967 [ptr target="#nyce1991" loc="57"/>. No other block of text remained unchanged over time; the technologies used to implement the mechanism changed, Memex grew \*intelligent\*, the other machines \(the Cyclops Camera, the Vocoder\) disappeared. These paragraphs, however, remain a constant. Given this fact, Nelson's assertion that the major concern of the essay was to point out the artificiality of systems of indexing, and to propose the associative mechanism as a solution for this \[ptr target="#nelson1991" loc="248"/> seems reasonable. Nelson also maintains that these central precepts of the design have been \\*quote rend="inline">ignored</i> by commentators \\[ptr target="#nelson1991" loc="245"/>. I would contend that they have not been \\\*\\\*ignored\\\*\\\*; fragments of these paragraphs are often cited, particularly relating to association. What is ignored is the relationship between these two paragraphs — the central \\\*\\\*contrast\\\*\\\* he makes between conventional methods of indexing and the mental associations Memex was to support \\\[ptr target="#nyce1991" loc="57"/>. Association was more \\\\*natural\\\\* than other forms of indexing — more human. This is why it was revolutionary.\\\]\\\(#\\\)</p>\\]\\(#\\)\\*\]\(#\)](#)

<p>Which is interesting, because Bush's model of mental association was itself technological; the mind *quote rend="inline">snapped</i> between allied items, an unconscious movement directed by the trails themselves, trails [cit>](#)*

*quote rend="inline">of brain or of machine</i>*

[ptr target="#bush1970" loc="191"/>](#)

</cit>. Association was a technique that worked independently of its substrate, and there was no spirit attached to this machine: [cit>](#)

*quote rend="inline">my brain runs rapidly — so rapidly I do not fully recognize that the process is going on</i>*

[ptr target="#bush1970" loc="191"/>](#)

</cit>. The *quote rend="inline">speed of action</i> in the retrieval process from neuron to neuron [ptr target="#bush1970" loc="102"/> resulted from a \*quote rend="inline">mechanical switching</i> \(this term was omitted from the \*title rend="italic">Life</i> reprint of \*title rend="italic">Memex II</i>, \[ref target="#bush1970">Bush 1970, 100</ref>\\), and the items that this mechanical\]\(#\)\*\*\*](#)*

process resurrected were also stored in the manner of magnetic or drum memory: the brain is like a substrate for [cit](#)

[quote](#) [rend="inline"](#)>memories, sheets of data[/quote](#)

[ptr](#) [target="#bush1970"](#) [loc="191"/>](#)

[/cit](#)>. [/p](#)

[p](#)>Bush's model of human associative memory was an electro-mechanical one — a model that was being keenly developed by Claude Shannon, Warren McCulloch and Walter Pitts at MIT, and would result in the McCulloch-Pitts neuron [ptr](#) [target="#hayles1999"](#)

[loc="65"/>](#). The MIT model of the human neuronal circuit constructed the human in

terms of the machine, and later articulated it more thoroughly in terms of computer switching. In a 1944 letter to Weeks, for example, Bush argued that [quote](#)

[rend="inline"](#)>a great deal of our brain cell activity is closely parallel to the operation of relay circuits[/quote](#), and that [cit](#)

[quote](#) [rend="inline"](#)>one can explore this parallelism...almost indefinitely[/quote](#)

[ref](#) [target="#nyce1991"](#) >November 6, 1944; cited in Nyce and Kahn 1991, 62[/ref](#)

[/cit](#)>. [/p](#)

[p](#)>In the 1930s and 1940s, the popular scientific conception of mind and memory was a mechanical one. An object or experience was perceived, transferred to the memory-library's receiving station, and then [cit](#)

[quote](#) [rend="inline"](#)>installed in the memory-library for all future reference[/quote](#)

[ptr](#) [target="#dennett1993"](#) [loc="121"/>](#)

[/cit](#)>. It had been known since the early 1900s that the brain comprised a tangle of neuronal groups that were interconnected in the manner of a network, and recent research had shown that these communicated and [q](#)stored[/q](#) information across the neural substrate, in some instances creating further connections, via minute electrical [q](#)vibrations[/q](#). According to Bush, memories that were not accessed regularly suffered from this neglect by the conscious mind and were prone to fade. The pathways of the brain, its indexing system, needed constant electrical stimulation to remain strong. This was the problem with the neural network: [cit](#)

[quote](#) [rend="inline"](#)>items are not fully permanent, memory is transitory[/quote](#)

[ptr](#) [target="#bush1945"](#) [loc="102"/>](#)

[/cit](#)>. The major technical problem with human memory was its tendency toward decay. [/p](#)

[p](#)>According to Manuel De Landa, there was also a widespread faith in biological-mechanical analogues at the time as models to boost human functions. The military had been attempting to develop technologies which mimicked and subsequently replaced human faculties for many years [ptr](#) [target="#delanda1994"](#) [loc="127"/>](#) and this was especially heightened in the years before, during and immediately following the war. At MIT in particular, there was a tendency to take [quote](#) [rend="inline"](#)>the image of the machine as the basis for the understanding of man[/quote](#) and vice versa, writes Harold Hatt in his book on Cybernetics [ptr](#) [target="#hatt1968"](#) [loc="28"/>](#). The idea that Man and his environment are mechanical systems which can be studied, improved, mimicked and controlled was growing, and later gave way to disciplines such as cognitive science and artificial intelligence. Wiener and McCulloch [cit](#)

[quote](#) [rend="inline"](#)>looked for and worked from parallels they saw between neural structure and process and computation[/quote](#)

[ptr](#) [target="#nyce1991"](#) [loc="63"/>](#)

[/cit](#)>, a model which changed with the onset of digital computing to include on/off states. The motor should first of all model itself on man, and eventually augment or replace him. [/p](#)

[p](#)>Bush explicitly worked with such methodologies — in fact, [cit](#)

[he not only thought with and in these terms, he built technological projects with them](#)

[#nyce1991 loc="62"](#)

. The first step was understanding the mechanical [process](#) or nature of thought itself; the second step was transferring this process to a machine. So there is a double movement within Bush's work, the location of a [natural](#) human process within thought, a process which is already machine-like, and the subsequent refinement and modelling of a particular technology on that process. Technology should depart from nature, it should depart from an extant human process: this saves us so much work. If this is done properly, [cit](#)

[\[it\] should be possible to beat the mind decisively in the permanence and clarity of the items resurrected from storage](#)

[#bush1970 loc="191"](#)

.

So Memex was first and foremost an extension of human memory and the associative movements that the mind makes through information: a mechanical analogue to an already mechanical model of memory. Bush transferred this idea into information management; Memex was distinct from traditional forms of indexing not so much in its mechanism or content, but in the way it organised information based on association. The design did not spring from the ether, however; the first Memex design incorporates the technical architecture of the Rapid Selector and the methodology of the Analyzer — the machines Bush was assembling at the time.

The Design of Memex

Bush's autobiography, [Pieces of the Action](#), and also his essay [Memex Revisited](#) tell us that he started work on the design in the early 1930s [#bush1967 loc="197"](#); [#bush1970 loc="130"](#). Nyce and Kahn also note that he sent a letter to Warren Weaver describing a Memex-like device in 1937 [#nyce1991 loc="43"](#). The first extensive description of it in print, however, is found in the 1939 essay [Mechanization and the Record](#)

[#bush1939](#). The description in this essay employs the same [methodology](#) Bush had used to design the Analyzer: combine existing lower-level technologies into a single machine with a higher function that automates the [pick-and-shovel](#) work of the human mind [#owens1991 loc="3"](#).

Nyce and Kahn maintain that Bush took this methodology from the Rapid Selector [#nyce1991 loc="54"](#): this paper has argued that it was first deployed in the Analyzer. The Analyzer was the first working analogue computer at MIT, and it was also the first large-scale engineering project to combine lower-level, extant technologies and automate what was previously a human cognitive technique: the integral calculus. It incorporated two lower-level analogue technologies to accomplish this task: the wheel-and-disk integrator and the torque amplifier, as we have explored. Surrounded by computers and personal organisers, the idea of automating intellectual processes seems obvious to us now — but in the early 1930s the idea of automating what was essentially a [function within thought](#) was radical. Bush needed to convince people that it was worthwhile. In 1939, Bush wrote:

[cit](#)

[The future means of implementing thought are ... fully as worthy of attention by one who wonders what comes next as are new ways of extracting natural resources, or of killing men.](#)

[<ptr target="#bush1939"/>](#bush1939)

</cit>

<p>The idea of creating a machine to aid the mind did not belong to Bush, nor did the technique of integral calculus (or association for that matter); he was, however, arguably the first person to externalise this technology on a grand scale. Observing the success of the Analyzer [qua](#) technical artefact, the method proved successful. Design on the first microfilm selection device, the Comparator, started in 1935. This, too, was a machine to aid the mind: it was essentially a counting machine, to tally the coincidence of letters in two messages or copies of a single message. It externalised the [drudge](#) work of cryptography, and Bush [cit](#)

[quote](#) [rightly saw it as the first electronic data-processing machine](#)

</cit>. The Rapid Selector which followed it incorporated much of the same architecture, as we have explored — and this architecture was in turn transferred to Memex. </p>

<cit>

[quote](#) [The Memex-like machine proposed in Bush's 1937 memo to Weaver shows just how much \[the Selector\] and the Memex have in common. In the rapid selector, low-level mechanisms for transporting 35mm film, photo-sensors to detect dot patterns, and precise timing mechanisms combined to support the high-order task of information selection. In Memex, photo-optic selection devices, keyboard controls, and dry photography would be combined ... to support the process of the human mind.](#)

</cit>

<p>The difference, of course, was that Bush's proposed Memex would access information stored on microfilm by [emph](#)association[emph](#), not numerical indexing. He had incorporated another technique (a technique which was itself quite popular among the nascent cybernetics community at MIT, and already articulated mind and machine together). By describing an imaginary machine, Bush had [cit](#)

[quote](#) [selected from the existing technologies of the time and made a case for how they should develop in the future](#)

</cit>. But this forecasting did not come from some genetically inherited genius — it was an acquired skill: Bush was close to the machine. </p>

<p>As Professor of Engineering at MIT (and after 1939, President of the Carnegie Institute in Washington), Bush was in a unique position — he had access to a pool of ideas, techniques and technologies which the general public, and engineers at other smaller schools, did not have access to. Bush had a more [q](#)global[q](#) view of the combinatorial possibilities and the technological lineage. Bush himself admitted this; in fact, he believed that engineers and scientists were the only people who could or [emph](#)should[emph](#) predict the future of technology — anyone else had no idea. In [title](#) [The Inscrutable Thirties](#)[title](#), an essay he published in 1933, he tells us that politicians and the general public simply can't understand technology, they have [cit](#)

[quote](#) [so little true discrimination](#)

</cit> and are [quote](#) [wont to visualize scientific triumphs as](#)

[foreign](#)faits accomplis[foreign](#)

[quote](#) before they are even ready, [cit](#)

[quote](#) [even as they are being hatched in the laboratory](#)



[bush1933](#bush1933) loc="75"/> Bush believed that the prediction and control of the future of technology should be left to engineers; only they can [cit](#) [quote rend="inline">distinguish the \[emph>possible</emph> from the virtually \\[emph>impossible</emph>\\]\\(#\\)</quote>\]\(#\)](#)

[nyce1991](#nyce1991) loc="49"/> [cit](#), only they can read the future from technical objects. </p>

Memex was a future technology. It was originally proposed as a desk at which the user could sit, equipped with two [quote rend="inline">slanting translucent screens</quote> upon which material would be projected for \[cit\]\(#\) \[quote rend="inline">convenient reading</quote>\]\(#\)](#)

[bush1945](#bush1945) loc="102"/> [cit](#). There was a keyboard to the right of these screens, and a [quote rend="inline">set of buttons and levers</quote> which the user could depress to search the information using an electrically-powered optical recognition system. If the user wished to consult a certain piece of information, \[cit\]\(#\) \[quote rend="inline">he \\[tapped\\] its code on the keyboard, and the title page of the book promptly appear\\[ed\\]</quote>\]\(#\)](#)

[bush1945](#bush1945) loc="103"/> [cit](#). The images were stored on microfilm inside the desk, [quote rend="inline">and the matter of bulk \[was\] well taken care of</quote> by this technology — \[cit\]\(#\) \[quote rend="inline">only a small part of the interior is devoted to storage, the rest to mechanism</quote>\]\(#\)](#)

[bush1945](#bush1945) loc="102"/> [cit](#). It looked like an [quote rend="inline">ordinary</quote> desk, except it had screens and a keyboard attached to it. To add new information to the microfilm file, a photographic copying plate was also provided on the desk, but most of the Memex contents would be \[cit\]\(#\) \[quote rend="inline">purchased on microfilm ready for insertion</quote>\]\(#\)](#)

[bush1945](#bush1945) loc="102"/> [cit](#). The user could classify material as it came in front of him using a teleautograph stylus, and register links between different pieces of information using this stylus. This was a piece of furniture from the future, to live in the home of a scientist or an engineer, to be used for research and information management. </p>

The 1945 Memex design also introduced the concept of [quote rend="inline">trails</quote>, a concept derived from work in neuronal storage-retrieval networks at the time, which was a method of connecting information by linking units together in a networked manner, similar to hypertext paths. The process of making trails was called \[quote rend="inline">trailblazing</quote>, and was based on a mechanical provision \\[cit\\]\\(#\\) \\[quote rend="inline">whereby any item may be caused at will to select immediately and automatically another</quote>\\]\\(#\\)\]\(#\)](#)

[bush1945](#bush1945) loc="107"/> [cit](#), just as though these items were being [cit](#) [quote rend="inline">gathered together from widely separated sources and bound together to form a new book</quote>](#)

[bush1945](#bush1945) loc="104"/> [cit](#). Electro-optical devices borrowed from the Rapid Selector used spinning rolls of microfilm, abstract codes and a mechanical selection-head inside the desk to find and create these links between documents. [cit](#) [quote rend="inline">This is the essential feature of the Memex. The process of](#)

tying two items together is the important thing</quote>

<ptr target="#bush1945" loc="103"/>

</cit>. Bush went so far as to suggest that in the future, there would be professional trailblazers who took pleasure in creating useful paths through the common record in such a fashion. </p>

<p>The Memex described in <title rend="italic">As We May Think</title> was to have permanent trails, and public encyclopaedias, colleague's trails and other information could all be joined and then permanently archived for later use. Unlike the trails of memory, they would never fade. In <title rend="italic">Memex Revisited</title>, however, an adaptive theme emerged whereby the trails were mutable and open to growth and change by Memex itself as it observed the owner's habits of association and extended upon these <ptr target="#bush1967" loc="213"/>. After a period of observation, Memex would be given instructions to search and build a new trail of thought, which it could do later <cit>

<quote rend="inline">even when the owner was not there</quote>

<ptr target="#bush1967" loc="213"/>

</cit>. This technique was in turn derived from Claude Shannon's experiments with feedback and machine learning, embodied in the <quote rend="inline">mechanical mouse</quote>; <cit>

<quote rend="block">A striking form of self adaptable machine is Shannon's mechanical mouse. Placed in a maze it runs along, butts its head into a wall, turns and tries again, and eventually muddles its way through. But, placed again at the entrance, it proceeds through without error making all the right turns.</quote>

<ptr target="#bush1959" loc="171"/>

</cit>

</p>

<p>In modern terminology, such a machine is called an intelligent <q>agent</q>, a concept we shall discuss later in this work. Technology has not yet reached Bush's vision for adaptive associative indexing <ptr target="#meyrowitz1991" loc="289"/>, although intelligent systems, whose parameters change in accordance with the user's experiences, come close. This is called machine learning. Andries van Dam also believes this to be the natural future of hypertext and associative retrieval systems

<ptr target="#vandam1999"/>.</p>

<p>In <title rend="italic">Memex II</title>, however, Bush not only proposed that the machine might learn from the human via what was effectively a cybernetic feedback loop — he proposed that the <emph>human might learn from the machine</emph>. As the human mind moulds the machine, so too the machine <quote rend="inline"

>remolds</quote> the human mind, it <cit>

<quote rend="inline">remolds the trails of the user's brain, as one lives and works in close interconnection with a machine</quote>

<ptr target="#bush1959" loc="178"/>

</cit>. <cit>

<quote rend="block">For the trails of the machine become duplicated in the brain of the user, vaguely as all human memory is vague, but with a concomitant emphasis by repetition, creation and discard ... as the cells of the brain become realigned and reconnected, better to utilize the massive explicit memory which is its servant.</quote>

<ptr target="#bush1959" loc="178"/>

</cit>

</p>

<p>This was in line with Bush's conception of technical machines as mechanical teachers in their own right. It was a <cit>

<quote rend="inline">proposal of an active symbiosis between machine and human memory</quote>

<ptr target="#nyce1991" loc="122"/>

</cit> which has been surprisingly ignored in contemporary readings of the design.

Nyce and Kahn pay it a full page of attention, and also Nelson, who has always read

Bush rather closely <ptr target="#nelson1999"/>. But aside from that, the full

development of this concept from Bush's work has been left to Doug Engelbart. </p>

<p>In our interview, Engelbart claimed it was Bush's concept of a <quote rend="inline">

>co-evolution</quote> between humans and machines, and also his conception of our

human <quote rend="inline">augmentation system</quote>, which inspired him <ptr

target="#engelbart1999"/>. Both Bush and Engelbart believe that our social

structures, our discourses and even our language can and should <cit>

<quote rend="inline">adapt to mechanization</quote>

<ptr target="#bush1967" loc="210"/>

</cit>; all of these things are inherited, they are learned. This process is not only

unavoidable, it is desirable. Bush also believed machines to have their own logic,

their own <emph>language</emph>, which <quote rend="inline">can touch those subtle

processes of mind, its logical and rational processes</quote> and alter them <ptr

target="#bush1959" loc="177"/>. And the <quote rend="inline">logical and rational

processes</quote> which the machine connected with were <emph>our own

memories</emph> — a prosthesis of the inside. This vision of actual human neurons

changing to be more like the machine, however, would not find its way into the 1967

essay <ptr target="#nyce1991" loc="122"/>. </p>

<p>Paradoxically, Bush also retreats on this close alignment of memory and machine. In

the later essays, he felt the need to demarcate a purely <q>human</q> realm of

thought from technics, a realm uncontaminated by technics. One of the major themes in

<title rend="italic">Memex II</title> is defining exactly what it is that machines

can and cannot do. </p>

<cit>

<quote rend="block">Two mental processes the machine can do well: first, memory

storage and recollection, and this is the primary function of the Memex; and

second, logical reasoning, which is the function of the computing and analytical

machines.</quote>

<ptr target="#bush1959" loc="178"/>

</cit>

<p>Machines can remember better than human beings can — their trails do not fade, their

logic is never flawed. Both of the <quote rend="inline">mental processes</quote> Bush

locates above take place within human thought, they are forms of internal <quote

rend="inline">repetitive</quote> thought <ptr target="#bush1967" loc="189"/> —

perfectly suited to being externalised and improved upon by technics. But exactly

what is it that machines <emph>can't </emph>do? Is there anything inside thought

which is purely human? Bush demarcates <quote rend="inline">creativity</quote> as the

realm of thought that exists beyond technology. </p>

<cit>

<quote rend="block">How far can the machine accompany and aid its master along this

path? Certainly to the point at which the master becomes an artist, reaching into

the unknown with beauty and versatility, erecting on the mundane thought processes

a thing of beauty ... this region will always be barred to the machine.</quote>

<ptr target="#bush1959" loc="183"/>

</cit>

<p>Bush had always been obsessed with memory and technics, as we have explored. But near

the end of his career, when <title rend="italic">Memex II</title> and <title

rend="italic">Memex Revisited</title> were written, he became obsessed with the



<quote rend="inline">boundary</quote> between them, between what is personal and belongs to the human alone, and what can be or <emph>already is</emph> automated within thought.</p>

<p>In all versions of the Memex essay, the machine was to serve as a personal memory support. It was not a <emph>public</emph> database in the sense of the modern Internet: it was first and foremost a private device. It provided for each person to add their own marginal notes and comments, recording reactions to and trails from others' texts, and adding selected information and the trails of others by <quote rend="inline">dropping</quote> them into their archive via an electro-optical scanning device. In the later adaptive Memex, these trails faded out if not used, and <cit> <quote rend="inline">if much in use, the trails become emphasized</quote> <ptr target="#bush1970" loc="191"/> </cit> as the web adjusts its shape mechanically to the thoughts of the individual who uses it. </p>

<p>Current hypertext technologies are not quite so private and tend to emphasise <cit> <quote rend="inline">systems which are public rather than personal in nature and that emphasize the static record over adaptivity</quote> <ptr target="#oren1991" loc="320"/>

</cit> due to the need for mass production, distribution and compatibility. The idea of a <q>personal</q> machine to amplify the mind also flew in the face of the emerging paradigm of human–computer interaction that reached its peak in the late 1950s and early 1960s, which held computers to be rarefied calculating machines used only by qualified technicians in white lab coats in air-conditioned rooms at many degrees of separation from the <q>user</q>. <quote rend="inline">After the summer of 1946</quote>, writes Ceruzzi, <cit>

<quote rend="inline">computing's path, in theory at least, was clear</quote> <ptr target="#ceruzzi1998" loc="23"/>

</cit>. Computers were, for the moment, impersonal, institutionally aligned and out of the reach of the ignorant masses who did not understand their workings. They lived only in university computer labs, wealthy corporations and government departments. Memex II was published at a time when the dominant paradigm of human–computer interaction was sanctified and imposed by corporations like IBM, and <cit>

<quote rend="inline">it was so entrenched that the very idea of a free interaction between users and machines as envisioned by Bush was viewed with hostility by the academic community</quote> <ptr target="#delanda1994" loc="219"/>

</cit>. </p>

<p>In all versions of the essay, Memex remained profoundly uninfluenced by the paradigm of digital computing. As we have explored, Bush transferred the concept of machine learning from Shannon — but not information theory. He transferred neural and memory models from the cybernetic community — but not digital computation. The analogue computing discourse Bush and Memex created never <quote rend="inline">mixed</quote> with digital computing <ptr target="#bush1970" loc="208"/>. In 1945, Memex was a direct analogy to Bush's conception of human memory; in 1967, after digital computing had swept engineering departments across the country into its paradigm, Memex was still a direct analogy to human memory. It mirrored the technique of association in <emph>its mechanical workings</emph>. <cit>

<quote rend="block">While the pioneers of digital computing understood that machines would soon accelerate human capabilities by doing massive calculations, Bush continued to be occupied with extending, through replication, human mental experience.</quote>

<ptr target="#nyce1991" loc="124"/> </cit>

</p>

<p>Consequently, the Memex redesigns responded to the advances of the day quite differently to how others were responding at the time. By 1967, for example, great advances had been made in digital memory techniques. As far back as 1951, the Eckert-Mauchly division of Remington Rand had turned over the first <q>digital</q> computer with a stored-program architecture, the UNIVAC, to the US Census Bureau <ptr target="#ceruzzi1998" loc="27"/>. <q>Delay Lines</q> stored 1,000 words as acoustic pulses in tubes of mercury, and reels of magnetic tapes which stored invisible bits were used for bulk memory. This was electronic digital technology, and did not mirror or seek to mirror <q>natural</q> processes in any way. It steadily replaced the most popular form of electro-mechanical memory from the late 1940s and early 1950s: drum memory. This was a large metal cylinder which rotated rapidly beneath a mechanical head, where information was written across the surface magnetically <ptr target="#ceruzzi1998" loc="38"/>. In 1957, disk memory had been produced, for the IBM305 RAMAC, and rapid advances were being made by IBM and DEC <ptr target="#ceruzzi1998" loc="196"/>. </p>

<p>Bush, however, remained enamoured of physical recording and inscription. His 1959 essay proposes using organic crystals to record data by means of phase changes in molecular alignment. <cit>

<quote rend="inline">[I]n Memex II, when a code on one item points to a second, the first part of the code will pick out a crystal, the next part the level in this, and the remainder the individual item</quote>

<ptr target="#bush1959" loc="169"/>

</cit>. This was new technology at the time, but certainly not the direction commercial computing was taking via DEC or IBM. Bush was fundamentally uncomfortable with digital electronics as a means to store material. <quote rend="inline">The brain does not operate by reducing everything to indices and computation</quote>, Bush wrote <ptr target="#bush1965" loc="190"/>. Bush was aware of how out of touch he was with emerging digital computing techniques, and this essay bears no trace of engineering details whatsoever, details which were steadily disappearing from all his published work. He devoted the latter part of his career to frank prophecy, reading from the technologies he saw around him and taking <cit>

<quote rend="inline">a long look ahead</quote>

<ptr target="#bush1959" loc="166"/>

</cit>. Of particular concern to him was promoting Memex as the technology of the future, and encouraging the public that <cit>

<quote rend="inline">the time has come to try it again</quote>

<ptr target="#bush1959" loc="166"/>

</cit>.</p>

</div>

<div>

<head>Memex, Inheritance and Transmission</head>

<quote rend="block">No memex could have been built when that article appeared. In the quarter-century since then, the idea has been with me almost constantly, and I have watched new developments in electronics, physics, chemistry and logic to see how they might help bring it to reality <ptr target="#bush1970" loc="190"/>.</quote>

<p>Memex became an image of potentiality for Bush near the end of his life. In the later essays, he writes in a different tone entirely: Memex was an image he would bequeath to the future, a gift to the human race. For most of his professional life, he had been concerned with augmenting human memory, and preserving information that might be lost to human beings. He had occasionally written about this project as a larger idea which would boost <quote rend="inline">the entire process by which man profits by his inheritance of acquired knowledge</quote>

[#bush1945](#) loc="99"/>. But in [Memex II](#), this project became grander, more urgent — the idea itself far more important than the technical details. He was nearing the end of his life, and Memex was still unbuilt. Would someone eventually build this machine? He hoped so, and he urged the public that it would soon be possible to do this, or at least, the [day has come far closer](#)

[#bush1970](#) loc="190"/>: [in the interval since that paper \[As We May Think\]](#) was published, there have been many developments ... steps that were merely dreams are coming into the realm of practicality

[#bush1959](#) loc="166"/>. Could this image be externalised now, and live beyond him? It would not only carry the wealth of his own knowledge beyond his death, it would be like a gift to all mankind. In fact, Memex would be the centrepiece of mankind's **true revolution** — transcending death.

**Can a son inherit the memex of his father, refined and polished over the years, and go on from there? In this way can we avoid some of the loss which comes when oxygen is no longer furnished to the brain of the great thinker, when all the patterns of neurons so painstakingly refined become merely a mass of protein and nucleic acid? Can the race thus develop leaders, of such power and intellect, and such forces of conviction, that the world can be saved from its follies? This is an objective of far greater importance than the conquest of disease, even than the conquest of mental aberrations**

Near the end of his life, Bush thought of Memex as more than just an

**individual's** machine; the

[ultimate \[machine\] is far more subtle than this](#)

[#bush1959](#) loc="182"/>

. Memex would be the centrepiece of a structure of inheritance and transmission, a structure that would accumulate with each successive generation. In

[Science Pauses](#), Bush entitled one of the sections

[Immortality in a machine](#)

[#bush1965](#) loc="189"/>: it contained a description of Memex, but this time there was an emphasis on its [longevity](#) over the individual human mind [#bush1965](#) loc="190"/>. This is the crux of the matter; the trails in Memex would not grow old, they would be a gift from father to son, from one generation to the next.

Bush died on June 30, 1974. The image of Memex has been passed on beyond his death, and it continues to inspire a host of new machines and [technical](#)

instrumentalities. But Memex itself has never been built; it exists only on paper, in technical interpretation and in memory. All we have of Memex are the words that Bush assembled around it in his lifetime, the drawings created by the artists from [Life](#), its erotic simulacrum, its ideals, its ideas. Had Bush attempted to assemble this machine in his own lifetime, it would undoubtedly have changed in its technical workings; the material limits of microfilm, of photoelectric components and later, of crystalline memory storage would have imposed their limits; the [use function](#) of the machine would itself have changed as it demonstrated its own potentials. If Memex had been built, the object would have invented itself independently of the outlines Bush cast on paper. This never happened — it has entered into the intellectual capital of new media as an image of potentiality.

<bibl xml:id="bagg1961" label="Bagg and Stevens 1961" key="bagg1961">  
 <author>Bagg, T. C.</author>, and <author>Stevens, M. E.</author>  
 <title rend="italic">Information Selection Systems Retrieving Replica Copies: A  
 state-of-the-art report</title>. <title rend="none">National Bureau of Standards  
 Technical note 157</title>. <pubPlace>Washington, D.C.</pubPlace>:  
 <publisher>Government Printing Office</publisher>, <date>1961</date>.</bibl>

<bibl xml:id="beniger1986" label="Beniger 1986" key="beniger1986">  
 <author>Beniger, James R.</author>  
 <title rend="italic">The Control Revolution: Technological and Economic Origins of  
 the Information Society</title>. <pubPlace>Cambridge, MA</pubPlace>:  
 <publisher>Harvard University Press</publisher>, <date>1986</date>.</bibl>

<bibl xml:id="burke1991" label="Burke 1991" key="burke1991">  
 <author>Burke, Collin</author>. <title rend="quotes">A Practical View of the Memex:  
 The Career of the Rapid Selector</title>. In <ptr target="#nyce1991"/>.</bibl>

<bibl xml:id="bush1928" label="Bush 1928" key="bush1928">  
 <author>Bush, Vannevar</author>. <title rend="quotes">Mechanical Solutions of  
 Engineering Problems</title>, <title rend="italic">Tech Engineering News</title>,  
 <biblScope type="vol">Vol. 9</biblScope>, <date>1928</date>.</bibl>

<bibl xml:id="bush1991" label="Bush 1991" key="bush1991">  
 <author>Bush, Vannevar</author>. <title rend="quotes">The Inscrutable <q>Thirties</q>  
 </title>. Reprinted in <ptr target="#nyce1991"/>, <biblScope type="pages">  
 >67–80</biblScope>.</bibl>

<bibl xml:id="bush1939" label="Bush 1939" key="bush1939">  
 <author>Bush, Vannevar</author>. <title rend="quotes">Mechanization and the  
 Record</title>, <title rend="italic">Vannevar Bush Papers</title>, Library of  
 Congress, Box 138, Speech Article Book File.</bibl>

<bibl xml:id="bush1945a" label="Bush 1945a" key="bush1945a">  
 <author>Bush, Vannevar</author>. <title rend="quotes">As We May Think</title>.  
 Reprinted in <ptr target="#nyce1991"/>, <biblScope type="pages">  
 >85–112</biblScope>.</bibl>

<bibl xml:id="bush1991a" label="Bush 1991a" key="bush1991a">  
 <author>Bush, Vannevar</author>. <title rend="quotes">Memex II</title>. Reprinted in  
 <ptr target="#nyce1991"/>, <biblScope type="pages">165–184</biblScope>.</bibl>

<bibl xml:id="bush1963" label="Bush 1963" key="bush1963">  
 <author>Bush, Vannevar</author>. <title rend="italic">Man's Thinking  
 Machines</title>, Vannevar Bush Papers, MIT Archives, MC78, Box 21.</bibl>

<bibl xml:id="bush1991d" label="Bush 1991d" key="bush1991d">  
 <author>Bush, Vannevar</author>. <title rend="quotes">Science Pauses</title>.  
 Reprinted in <ptr target="#nyce1991"/>, <biblScope type="pages">  
 >185–196</biblScope>.</bibl>

<bibl xml:id="bush1991b" label="Bush 1991b" key="bush1991b">  
 <author>Bush, Vannevar</author>. <title rend="quotes">Memex Revisited</title>.  
 Reprinted in <ptr target="#nyce1991"/>, <biblScope type="pages">  
 >197–216</biblScope>.</bibl>

<bibl xml:id="bush1970" label="Bush 1970" key="bush1970">  
 <author>Bush, Vannevar</author>. <title rend="italic">Pieces of the Action</title>.  
 <pubPlace>New York</pubPlace>: <publisher>William Morrow</publisher>,  
 <date>1970</date>.</bibl>

<bibl xml:id="ceruzzi1998" label="Ceruzzi 1998" key="ceruzzi1998">  
 <author>Ceruzzi, Paul E.</author>  
 <title rend="italic">A History of Modern Computing</title>. <pubPlace>Cambridge,  
 MA</pubPlace>: <publisher>MIT Press</publisher>, <date>1998</date>.</bibl>

<bibl xml:id="delanda1994" label="De Landa 1994" key="delanda1994">



<author>De Landa, Manuel</author>. <title rend="italic">War in the Age of Intelligent Machines</title>. <pubPlace>New York</pubPlace>: <publisher>Zone Books</publisher>, <date>1994</date>.</bibl>

<bibl xml:id="dennett1993" label="Dennett 1993" key="dennett1993">  
 <author>Dennett, Daniel C.</author>  
 <title rend="italic">Consciousness Explained</title>. <pubPlace>London</pubPlace>:  
 <publisher>Penguin Books</publisher>, <date>1993</date>.</bibl>

<bibl xml:id="edwards1997" label="Edwards 1997" key="edwards1997">  
 <author>Edwards, Paul N.</author>  
 <title rend="italic">The Closed World: Computers and the Politics of Discourse in  
 Cold War America</title>. <pubPlace>Cambridge, MA</pubPlace>: <publisher>MIT  
 Press</publisher>, <date>1997</date>.</bibl>

<bibl xml:id="eldredge2004" label="Eldredge 2004" key="[unlisted]">  
 <author>Eldredge, Niles.</author>  
 <title rend="none">Email interview with Belinda Barnet</title>. <date>March  
 2004</date>. <ref  
 target="http://journal.fibreculture.org/issue3/issue3\_barnet.html"  
 >http://journal.fibreculture.org/issue3/issue3\_barnet.html</ref>.</bibl>

<bibl xml:id="engelbart1999" label="Engelbart 1999" key="[unlisted]">  
 <author>Engelbart, Douglas</author>. <title rend="none">Interview with Belinda  
 Barnet</title>. <date>November 10, 1999</date>.</bibl>

<bibl xml:id="farkas1990" label="Farkas1990" key="farkas1990">  
 <author>Farkas-Conn, I. S.</author>  
 <title rend="italic">From Documentation to Information Science: The Beginnings and  
 Early Development of the American Documentation Institute—American Society for  
 Information Science</title>. <pubPlace>New York</pubPlace>: <publisher>Greenwood  
 Press</publisher>, <date>1990</date>.</bibl>

<bibl xml:id="guattari1995" label="Guattari 1995" key="guattari1995">  
 <author>Guattari, Félix</author>. <title rend="italic">Chaosmosis: An  
 Ethico-Aesthetic Paradigm</title>. Tr. <name role="translator">Paul Bains</name>  
 and <name role="translator">Julian Pefanis</name>. <pubPlace>Sydney</pubPlace>:  
 <pubPlace>Power Publications</pubPlace>, <date>1995</date>.</bibl>

<bibl xml:id="gille1986" label="Gille 1986" key="gille1986">  
 <author>Gille, Bertrand</author>. <title rend="italic">History of Techniques</title>.  
 <pubPlace>New York</pubPlace>: <publisher>Gordon and Breach Science  
 Publishers</publisher>, <date>1986</date>.</bibl>

<bibl xml:id="hartree2000" label="Hartree 2000" key="hartree2000">  
 <author>Hartree, Douglas</author>. <title rend="quotes">Differential  
 Analyzer</title>, <ref  
 target="http://cs.union.edu/~hemmendd/encyc/articles/difanal/difanal.html"  
 >http://cs.union.edu/~hemmendd/Encyc/Articles/Difanal/difanal.html</ref>  
 </bibl>

<bibl xml:id="hatt1968" label="Hatt 1968" key="hatt1968">  
 <author>Hatt, Harold</author>. <title rend="italic">Cybernetics and the Image of  
 Man</title>. <pubPlace>Nashville</pubPlace>: <publisher>Abingdon  
 Press</publisher>, <date>1968</date>.</bibl>

<bibl xml:id="hayles1993" label="Hayles 1993" key="hayles1993">  
 <author>Hayles, Katherine</author>. <title rend="quotes">Virtual Bodies and  
 Flickering Signifiers</title>, <title rend="italic">October Magazine</title>.  
 <biblScope type="vol">No 66</biblScope> (<date>Fall 1993</date>), <biblScope  
 type="pages">69–91</biblScope>.</bibl>

<bibl xml:id="hayles1999" label="Hayles 1999" key="hayles1999">  
 <author>Hayles, N. Katherine</author>. <title rend="italic">How we Became

Posthuman</title>. <pubPlace>Chicago</pubPlace>: <publisher>University of Chicago Press</publisher>, <date>1999</date>.</bibl>

<bibl xml:id="hazen1940" label="Hazen 1940" key="hazen1940">  
 <author>Hazen, Harold</author>. <title rend="none">MIT President's report</title>,  
 <date>1940</date>.</bibl>

<bibl xml:id="meyrowitz1991" label="Meyrowitz 1991" key="meyrowitz1991">  
 <author>Meyrowitz, Norman</author>. <title rend="quotes">Hypertext: Does it Reduce  
 Cholesterol, Too?</title>. In <ptr target="#nyce1991"/>, <biblScope type="pages">  
 >287–318</biblScope>.</bibl>

<bibl xml:id="mindell2000" label="Mindell 2000" key="mindell2000">  
 <author>Mindell, David A.</author>  
 <title rend="quotes">MIT Differential Analyzer</title>. <ref  
 target="http://web.mit.edu/mindell/www/analyzer.htm">  
 http://web.mit.edu/mindell/www/analyzer.htm</ref>  
 </bibl>

<bibl xml:id="nelson1991" label="Nelson 1991" key="nelson1991">  
 <author>Nelson, Theodor H.</author>  
 <title rend="quotes">As We Will Think</title>. In <ptr target="#nyce1991"/>,  
 <biblScope type="pages">245–260</biblScope>.</bibl>

<bibl xml:id="nelson1999" label="Nelson 1999" key="[unlisted]">  
 <author>Nelson, Theodor H.</author> Interview with the author.</bibl>

<bibl xml:id="nyce1991" label="Nyce 1991" key="nyce1991">  
 <editor>Nyce, James</editor> and <editor>Kahn, Paul</editor>, eds. <title  
 rend="italic">From Memex to Hypertext: Vannevar Bush and the Mind's  
 Machine</title>. <pubPlace>London</pubPlace>: <publisher>Academic  
 Press</publisher>, <date>1991</date>.</bibl>

<bibl xml:id="oren1991" label="Oren 1991" key="oren1991">Oren, Tim 1991, <title  
 rend="quotes">Memex: Getting Back on the Trail</title>. In <ptr target="#nyce1991"/>,  
 <biblScope type="pages">319–338</biblScope>.</bibl>

<bibl xml:id="owens1991" label="Owens 1991" key="owens1991">  
 <author>Owens, Larry</author>. <title rend="quotes">Vannevar Bush and the  
 Differential Analyzer: The Text and Context of an Early Computer</title>. In <ptr  
 target="#nyce1991"/>, <biblScope type="pages">3–38</biblScope>.</bibl>

<bibl xml:id="shurkin1996" label="Shurkin 1996" key="shurkin1996">  
 <author>Shurkin, Joel</author>. <title rend="italic">Engines of the Mind, The  
 Evolution of the Computer from Mainframes to Microprocessors</title>.  
 <pubPlace>New York</pubPlace>: <publisher>WW Norton and Company</publisher>,  
 <date>1996</date>.</bibl>

<bibl xml:id="smith1991" label="Smith 1991" key="smith1991">  
 <author>Smith, Linda C.</author>  
 <title rend="quotes">Memex as an Image of Potentiality Revisited</title>. In <ptr  
 target="#nyce1991" loc="261–286"/>.</bibl>

<bibl xml:id="spar2001" label="Spar 2001" key="spar2001">  
 <author>Spar, Debora L.</author>  
 <title rend="italic">Ruling the Waves: Cycles of Discovery, Chaos, and Wealth from  
 Compass to the Internet</title>. <pubPlace>New York</pubPlace>:  
 <publisher>Harcourt</publisher>, <date>2001</date>.</bibl>

<bibl xml:id="stiegler1998" label="Stiegler 1998" key="stiegler1998">  
 <author>Stiegler, Bernard</author>. <title rend="italic">Technics and Time, 1: The  
 Fault of Epimetheus</title>. <pubPlace>Stanford</pubPlace>: <publisher>Stanford  
 University Press</publisher>, <date>1998</date>.</bibl>

<bibl xml:id="vandam1999" label="Van Dam 1999" key="[unlisted]">  
 <author>Van Dam, Andries</author>. Interview with the author.</bibl>

<bibl xml:id="weaver1950" label="Weaver 1950" key="[unlisted]">  
 <author>Weaver, Warren</author>. <title rend="none">Project diaries</title>.  
 <date>March 17, 1950</date>.</bibl>  
<bibl xml:id="weaver1950a" label="Weaver 1950a" key="weaver1950a">  
 <author>Weaver, Warren</author>. <title rend="none">Letter to Samuel  
 Caldwell</title>. Correspondence held in the Rockefeller Archive Center, <idno  
 type="shelfmark">RF1.1/224/2/26</idno>.</bibl>  
<bibl xml:id="wells1938" label="Wells 1938" key="wells1938">  
 <author>Wells, H.G.</author>  
 <title>World Brain</title>. <pubPlace>London</pubPlace>: <publisher>Methuen & Co.  
 Limited</publisher>, <date>1938</date>.</bibl>  
<bibl xml:id="ziman2003" label="Ziman 2003" key="ziman2003">  
 <author>Ziman, John</author>. <title rend="italic">Technological Innovation as an  
 Evolutionary Process</title>. <pubPlace>Cambridge</pubPlace>: <publisher>Cambridge  
 University Press</publisher>, <date>2003</date>.</bibl>

</listBibl>

</back>

</text>

</TEI>

### ***Appendix 3: Side-by-side Comparison Layout***

This appendix contains a screen shot of the side-by-side comparison view used to identify mismatched bibliographic entries during the deduplication and error correction phase of the project. This view takes data from the bibliography for an individual DHQ article; for each entry in that bibliography, the XSLT stylesheet seeks a match (based on the value of the @key attribute) in the centralized bibliography. If a match is found, that entry is displayed beneath the original entry. The stylesheet also performs a comparison between the content of the two entries (based on author name, title, and facts of publication); if the similarity falls below a certain threshold, the entry is flagged in red so that the two can be compared and the match confirmed. In the examples shown here, the first flagged entry (Borovoy) is in fact a match but there are discrepancies between the titles; the entry from the central bibliography contains better information. In the second flagged entry (Marino) the two records represent different items and the @key will need to be fixed to point to the correct entry in the central bibliography.



# Code as Ritualized Poetry: The Tactics of the Transborder Immigrant Tool

Show Key

Show Instructions

Comparing with 6239 entries in Biblio.

//bibl[@key='bakhtin1982']

Bakhtin, M. M. The Dialogic Imagination: Four Essays. University of Texas Press, 1982.

Bakhtin, M.M. *The Dialogic Imagination: Four Essays*. Austin: University of Texas Press, 1982.

[Biblio also has 2 similar entries]

Show Detail

//bibl[@key='borovoy2011']

Borovoy, Rick et al. Folk Computing. ACM Press, 2001. 466–473. Web. 1 Oct. 2011.

Borovoy, Rick, et al. “Folk Computing”. Presented at (2001). <http://dl.acm.org/citation.cfm?id=365316>.

Hide Detail

Biblio ID	criterion	Biblio entry
borovoy2011	Matching ID Similarity 0.231 (6/26)	Borovoy, Rick, et al. “Folk Computing”. Presented at (2001). <a href="http://dl.acm.org/citation.cfm?id=365316">http://dl.acm.org/citation.cfm?id=365316</a> .

//bibl[@key='marino']

Brett Stalbaum Complete Interview : Mark Marino : Free Download & Streaming : Internet Archive. Film.

Marino, Mark. *Brett Stalbaum Complete Interview : Mark Marino*. Internet Archive.  
<https://archive.org/details/BrettStalbaumCompleteInterview>.

Show Detail

//bibl[@key='camnitzer2007']

Camnitzer, Luis. Conceptualism in Latin American Art: Didactics of Liberation. 1st ed. Austin: University of Texas Press, 2007. Print. (Joe R. and Teresa Lozano Long Series in Latin

#### ***Appendix 4: Internal Documentation for Extraction of Bibliography Entries***

This appendix contains the internal documentation describing the process by which bibliographic data is extracted from existing DHQ articles and converted to the DHQ bibliographic markup.



# Biblio Workflow Instructions

0. Open the Biblio.xpr file in Oxygen so that you have access to the "project" materials.

1. Make sure you're using the most up-to-date version of DHQ's files (via SVN).

2. Open the .xml version of the article you are working on in Oxygen

3. If the article has no bibliography, move on to the next article in the workflow.

4. If there are bibl records, extract them from the article; these records (after you de-duplicate and groom them) will become part of the Biblio list:

Configure Transformation Scenario (wrench icon next to the red "Apply Transformation Scenario" arrow)

Click the check-box next to "Extract biblio listings," then "Apply Associated (1)."

A new file, titled "numberoffile-biblioscratch.xml," should be created.

5. **IMPORTANT: Run a Find/Replace on the biblioscratch.xml file to convert all references to "dhqID" (an old referent) to "ID."**

- go to the Find menu and choose Find/Replace
- in Text to Find type: "dhqID", and in Replace With type "ID."
- click "Replace All"
- the number of matches should equal the number of biblio records (for example, "88 records matched").

6. Next you're going to check for duplicate records: i.e., records that have already been entered by Jim / DHQ into the current repository of bibliographic records (visible in the "current" sub-folder in the "data" folder in DHQ). This is done by running a Schematron check which compares the contents of your scratch file to the existing contents of Biblio. The goal here is to eliminate from your scratch file any records that are already in Biblio.

You do NOT have to clean up any records that are already present in "current", and you can delete them from your scratch file without worrying that they will be disconnected from the article (which is why we're doing this in a "scratch" file).

- go to the "Validate" check-box at the top of Oxygen and open the drop-down menu by clicking the arrow next to it:
- choose "Validate With"
- if you do not see options visible here, find the dhqBiblio schema file in your working copy ( dhq/trunk/biblio/DHQ-Biblio-v2/schema/dhqBiblio-checkup.sch), then click "OK." (make sure you're using the checkup file here!)
- you should then receive a number of error messages in the "Errors" section of Oxygen.
- check the red exclamation points first; they provide the most accurate information re: bibliographic information that already resides in "current."
- then check the yellow exclamation points; they represent possible duplicates based on matching titles (but since titles are often the same, e.g. "Introduction", this isn't always indicative of a duplicate).

## Red Error Messages

When checking these exclamation points:

- go to the Biblio record noted in the error message (for example: dhqID 'aarseth1997' is already assigned to another entry; see Biblio-A.xml (aarseth1997).
- you can find these alphabetic files in the "current" folder.
- check to ensure that both entries are the same. You should also verify that the information in "current" is the most comprehensive: for example, if you notice that the author's full name is not in "Biblio-A," then please update that in "current."
- if the entry is the same, you can delete the entry *in your scratch file*.
- in some cases you'll find that while an ID has already been assigned, the entry in your article is different. After double-checking to ensure that information on both citations is accurate, you may need to assign the citation tied to your article a new entry. For example, if your 'aarseth1997' is different from the 'aarseth1997' in the "current" folder's A file, you should rename your entry 'aarseth1997a' (or b, if an "aarseth1997a" already exists, etc.). This issue pops up with the particularly prolific writers cited by DHQ's authors (McGann, Hayles, Flanders).
- check every red exclamation point in your error messages until you are satisfied that records are duplicates / resolved.

## Yellow Error Messages

- these error messages generally refer to titles that are similar to entries in the "current" folder. Compare these messages to the specified bibliographic files and determine if you're dealing with a duplicate or a new entry.
- in some cases, these error messages contain information that you've hopefully already resolved while going through the red exclamation points. However, there will inevitably be occasions when a duplicate title is present in an entry that we want to add to our records: different / revised editions of publications, generic titles that happen to overlap (like "Digital Media"), generic titles like "Wikipedia."
- in some cases you might find that the title listed in your scratch file could be revised (expanded to contain more information, changed because it is incorrect). Feel free to do so, but if you've otherwise established that you're dealing with the correct title and a new entry, then don't worry about the error message if it persists.

### **8. Update and encode the bibliographic records remaining in your scratch file to create a valid file in accordance with the Biblio schema (adding elements and attributes as needed to represent the various components of the bibliographic record).**

See the information about Bibliographic Elements on this page to determine what element to use for each item

These are my (Jim's) suggestions for how to quickly complete this work; feel free to do what is best for you, so long as the end result is the same (a clean file that we can add to DHQ's records).

- put Boilerplate content in scratch file
- change *every* record's Biblioltem element to an appropriate genre (e.g. <BookSection> and clean up information about authors and editors
- update additional information for each record by Biblioltem (start with books, then journal articles, then websites, etc.). You can use the find tool to jump from item to item and work more quickly through the file this way. I tend to start with JournalArticle records, since they involve adding the most information.
- clean up the entire file until it is valid
- any items that don't conform to an existing Biblio genre should be added to the Problem Genres file

#### *Boilerplate:*

I tend to dump the following text into the top and/or bottom of my scratch file, since I know I'll end up using them a lot and I'll want to paste this content into many records:

For all records with authors (i.e. most of them):

```
<author>
  <givenName></givenName>
  <familyName></familyName>
</author>
```

For Books:

```
<place></place>
<publisher></publisher>
<date></date>
```

For Journal Articles:

```
<journal issuance="continuing">
  <title></title>
  <volume></volume>
  <issue></issue>
  <date></date>
</journal>
<startingPage></startingPage>
<endingPage></endingPage>
```

#### *Issuances*

Information about issuance accompanies information about each Biblioltem; this information designates whether an item is "monographic" or "continuing"

monographic: Book, BookInSeries, ConferencePaper JournalArticle, Thesis, VideoGame

continuing: BlogEntry, book (when part of BookInSeries information), journal, WebSite

*Tips for Author information*

- whenever possible, use full names instead of initials for givenName information.
- use CorporateName for corporate authors (institutional entities, companies). CorporateName is most frequently used for WebSites where authors are unspecified
- If no author name is present and a CorporateName can not be determined, use the FullName field and write "Author Unknown."

**9. Make sure entire file is clean and valid and that your work has been updated via Subversion (i.e. COMMIT your changes).**

**10. Notify Julia and we have Wendell propagate the resulting Biblio records into the Biblio data.**

## ***Appendix 5: Final Report on IVMOOC Project***

This appendix contains the final report by members of the IVMOOC working group describing their analysis of the DHQ bibliographic data and presenting the resulting visualizations.

# Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly

Dulce Maria de la Cruz, Jake Kaupp, Max Kemman, Kristin Lewis, Teh-Hen Yu

**Abstract**—Digital Humanities Quarterly (DHQ) is a young journal that covers the intersection of digital media and traditional humanities. In this paper, we explore the publication patterns in DHQ through visualizations of co-authorship and bibliographic coupling networks in order to understand the cultures the journal represents. We find that DHQ consists largely of sole-authored papers (66%) and the authorship is dominated (75%) by authors publishing from North American institutions. Through the backbone of DHQ's bibliographic coupling network, we identify several communities of articles published in DHQ, and we analyze their collective abstracts using term frequency-inverse document frequency (TF-IDF) analysis. The extracted terms show that DHQ has wide coverage across the digital humanities, and that sub areas of DHQ can be identified through their citation behavior.

**Index Terms**—Digital Humanities, Information Visualization, Co-author network, Bibliographic Coupling, big tent

## INTRODUCTION

Digital Humanities (DH) is a field of research difficult to define due to its heterogeneity<sup>1</sup>. With its inclusionary ambitions, DH is regularly referred to as a 'big tent' [1] encompassing scholars from a wide variety of disciplines such as history, literature, linguistics, but also disciplines such as human-computer interaction and computer science. This collaborative, multidisciplinary approach to digital media makes DH an interesting field, but also difficult to grasp. A question is to what extent the big tent of DH represents a single, or actually a variety of cultures [1, 2].

The Digital Humanities Quarterly (DHQ) journal is arguably one of the largest journals aimed specifically at DH research, and covers all aspects of digital media in the humanities, representing a meeting point between digital humanities research and the wider humanities community [3]. Articles published in DHQ involve authors of multiple countries, institutions and disciplines who work on several subjects and areas related to digital media research. Under a recent grant from NEH (National Endowment for Humanities), DHQ has developed a centralized bibliography which supports the bibliographic referencing for the journal. To gain an understanding of the diversity of culture(s) in the DH, we are interested in how unique disciplinary cultures are represented in DHQ. Considering cultures are self-referential systems, we might expect that scholars from a certain culture are more likely to cite scholars from their own culture rather than from others [2]. As such, we expect citation behaviour to reflect disciplinary cultural norms. Therefore, visualizing and analysing the bibliographic data of DHQ not only

gives insights into the specific bibliographies from DHQ, it might give insight into the way the different epistemic cultures in the DH big tent interact with one another, and how this interaction and collaboration impacts the networks over time.

This paper reports on a project undertaken in the Information Visualization MOOC from Indiana University<sup>2</sup>. We have analysed the DHQ bibliographic data and created visualizations in order to discuss the following questions provided by the DHQ editors:

1. how citations reflect differences in academic culture at the institutional and geographic level
2. the changes to that culture over time.
3. correlations between article topics (reflected in keywords) and citation patterns

## 1 METHOD

### 1.1 Data

Two tables were extracted from the Client dataset:

1. dhq\_articles (178 records)
2. works\_cited\_in\_dhq (3823 records)

The attributes for both tables are: article id, authors, year, title, journal/conference/collection, abstract, cited references, and isDHQ.

The raw dataset posed several problems, including:

- missing articles,
- duplicate authors,
- double affiliations and inconsistencies,
- duplicated articles and citation self-loops,
- special characters, and
- incomplete information (lack of information regarding affiliation and country for each DHQ paper, and disciplines for authors).

The DHQ website<sup>3</sup> was therefore scraped using the tool Import.io<sup>4</sup> to find missing articles and to obtain information about affiliations for each author. Once that information was known, it was used to obtain the country associated with each institution by searching in the web. Custom programs in the R language were then used to create paper IDs (cite me as) similar to those used for the references and to

- 
- Dulce Maria de la Cruz is Freelance Data Analyst. E-mail: Dulce.Maria.delaCruz@gmail.com.
  - Jake Kaupp is Engineering Education Researcher in Queen's University, Canada. E-mail: jkaupp@gmail.com.
  - Max Kemman is PhD Candidate in University of Luxembourg, Luxembourg. E-mail: maxkemman@gmail.com.
  - Kristin Lewis is Science & Technology Policy Fellow at AAAS. E-mail: kristin.l.m.lewis@gmail.com.
  - Teh-Hen Yu is IT Professional. E-mail: tehhenyu@hotmail.com.

---

<sup>1</sup> See e.g. <http://whatisdigitalhumanities.com> for a wide variety of definitions from different scholars

<sup>2</sup> <http://ivmooc.cns.iu.edu/>

<sup>3</sup> <http://www.digitalhumanities.org/dhq/>

<sup>4</sup> <https://www.import.io/>

calculate the number of times each DHQ paper has been cited (times cited) and the number of references cited by each DHQ paper (count cited references). Furthermore, we assigned a discipline to each paper based on the first author's departmental affiliation as described in [4]. In order to produce a more detailed list of disciplinary culture, departmental affiliation was manually mapped to web of science subject areas. This information was eventually not used for the final visualizations, but left in the dataset for further exploration by others.

After validations, data mining/scraping, data processing with custom programs coding and a lot of manual work, we have come up with a master dataset with additional info added (cite me as, times cited, affiliation, country, count cited references, geocode, discipline, affiliations including departments info, and community, plus the keywords provided by editors of DHQ). To provide sufficient resolution, and categorical variables, for visualizations an author look-up table was created which contained the additional information outlined above but for each separate author for each article ID. Both the master datafile and the author lookup table are our primary sources of data to load for visualization and analysis.

The source code, final datasets, and resulting visualizations are available through github<sup>5</sup>.

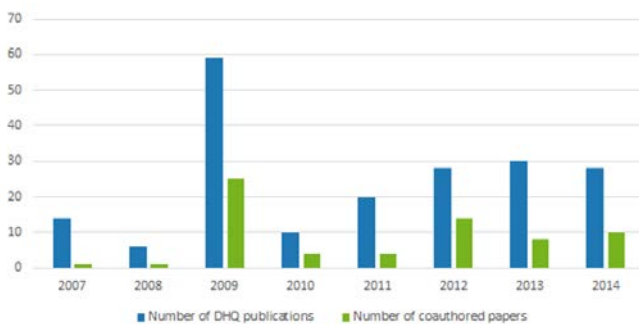
The final dataset provides the following statistics as in Table 1.

Table 1. DHQ dataset statistics

Attribute	Count	Note
DHQ articles	195	
Unique cited articles	4718	
Unique DHQ author	276	
Affiliations	148	Including all institutions + independent scholars
WOS subject areas	29	
Countries	17	
Publication years	8	2007-2014

Figure 1 provides an overview of the number of DHQ publications and number of co-authored papers per year, revealing a surprisingly uneven temporal distribution.

Fig. 1. DHQ (co-authored) publications per year.



## 1.2 Co-author network

<sup>5</sup> Available at <https://jkaupp.github.io/DHQ>. Please cite as Kaupp, J., De la Cruz, D.M., Kemman, M., Lewis, K., Yu, T.-H. (2015) Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly. GitHub, <https://jkaupp.github.io/DHQ>

People are the key inputs in determining and understanding cultural differences. Therefore, in order to better understand the cultures within DHQ, we explored the authors who published within DHQ. Using Sci2 [5], we created yearly cumulative time slices of the master dataset and extracted co-author networks for each time slice. Columns for author country were added, and each time slice was imported into Gephi to create a dynamic co-author network [6]. The network was laid out using the Force Atlas 2 algorithm [7], with nodes colorized by country. Each time slice was visualized, and compiled into comprehensive visualizations using Adobe Illustrator and Adobe Photoshop.

In addition to a co-author network, we explored a bibliographic coupling network of authors, in which nodes (authors) would be linked based on the number of cited articles in common. This analysis however introduced a strong bias towards co-authors who cite large numbers of articles. In order to derive useful insights from this type of visualization, a de-biasing operation must be identified and applied. Without an established method for these, we chose to focus on the geographic information in the co-authorship network and analyse bibliographic coupling of articles

## 1.3 Bibliographic coupling & Backbone identification

In order to investigate the bibliographies of DHQ articles, we analysed the data using Sci2 by extracting the paper-citation network, followed by extracting the reference co-occurrence network, also known as “bibliographic coupling” [8]. By doing so, we create a network of DHQ articles with co-occurring references. To simplify the visualization, we created a minimum spanning tree using the MST Pathfinder algorithm whereby articles are connected to the network only by their strongest relation [9], also called Backbone identification. As such, the network becomes a tree that is easier to read. Finally, all articles with zero references were removed from the network in order to remove non-DHQ articles, as well as DHQ articles that could not be analysed due to a lack of references.

This network was then analyzed using the SLM community detection algorithm with undirected and weighted edges [10]. The network with community attributes was then imported into Gephi and ordered using the Force Atlas 2 algorithm [6], after which we colorized the nodes by their identified community.

## 1.4 Word clouds

In order to investigate the correlations between article topics (reflected in keywords) and the citation patterns, word clouds of keywords were obtained for each of the communities identified via SLM detection in the bibliographic coupling network. For this purpose, community-based abstracts were obtained by combining the abstracts associated with the DHQ papers belonging to each community. These community-wide abstracts were normalized to lower case, tokenized, and stop words were removed. Words were not stemmed in order to differentiate between words like digital and digitized. Unique keywords were extracted from the community-based abstracts with custom R programs (using the R packages stringr<sup>6</sup> and tm<sup>7</sup>). The most significant keywords for each community were then identified through the Term frequency - Inverse Document Frequency (TF-IDF) method [11]. Terms with high TF-IDF values imply a strong relationship with the document in which they appear.

In this specific case, the terms are the unique keywords and the corpus of documents is the set of community-based abstracts. Therefore, the higher the TF-IDF value of a keyword in a

<sup>6</sup> <http://cran.r-project.org/web/packages/stringr/index.html>

<sup>7</sup> <http://cran.r-project.org/web/packages/tm/index.html>



The ten top-scoring words from each community were put into a word cloud and the words were sized by TF-IDF score. The word clouds were manually adjusted to unify the appearance of terms (plural vs. singular, infinitive vs. gerund, etc.) and were added to the bibliographic coupling network visualization.

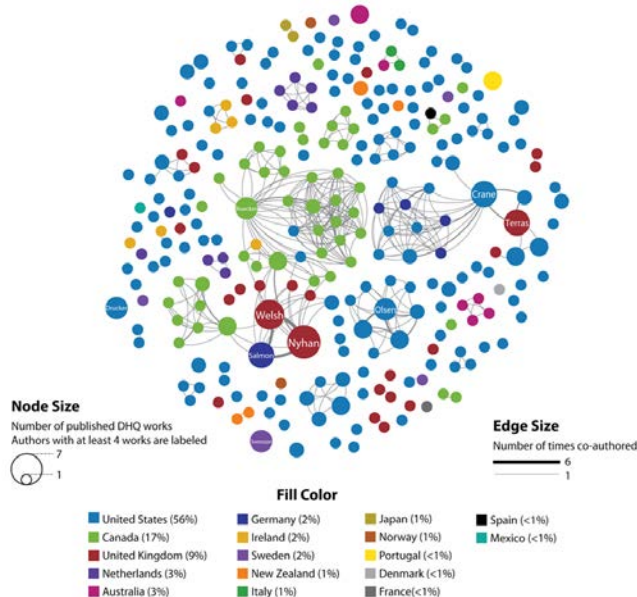
## 2 RESULT<sup>8</sup>

## 2.1 Co-author Network

Figures 2 and 3 represent the co-author network for DHQ, both comprehensively (Figure 2) and through cumulative time slices (Figure 3). Nodes are sized by the number of works published in DHQ, and in Figure 2, authors with at least 4 DHQ publications are labeled with the author's last name. Nodes are colored by the country of the author. The edges are weighted by the number of times each pair co-

Fig. 2 Co-Author network, 2007-2014

### Co-Author Network for Digital Humanities Quarterly: 2007-2014



authored a DHQ publication together. The maximum number of authored works (articles) for a single author is 7: Julianne Nyhan from UK. The maximum co-authored articles for two authors are 6: by Anne Welsh and Julianne Nyhan from UK. The most active year is 2009, as also shown in Figure 1, with several authors publishing multiple papers in this year.

## 2.2 Bibliographic coupling network with word clouds

Figure 4 shows the backbone bibliographic coupling network for DHQ, representing the strongest connections in the larger bibliographic coupling network (not shown). Nodes are colored by community, as identified through SLM detection, and sized by the number of articles cited in each article. Edges are weighted by the number of cited articles in common. Alongside each community is a

<sup>8</sup> Larger versions of all visualizations are available in the github repository.

word cloud of keywords in the same color extracted from the abstracts of each article in the community.

### Co-Author Network for Digital Humanities Quarterly: 2007-2014

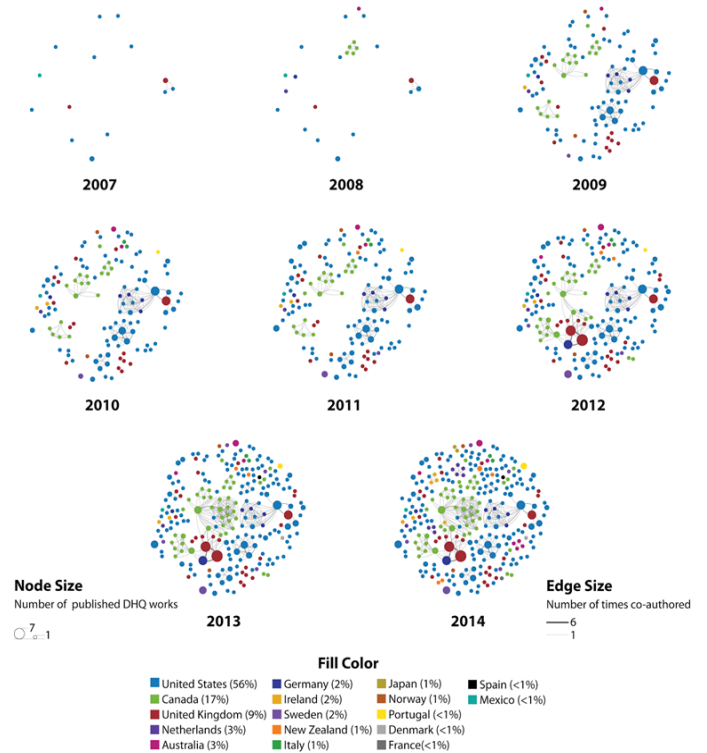


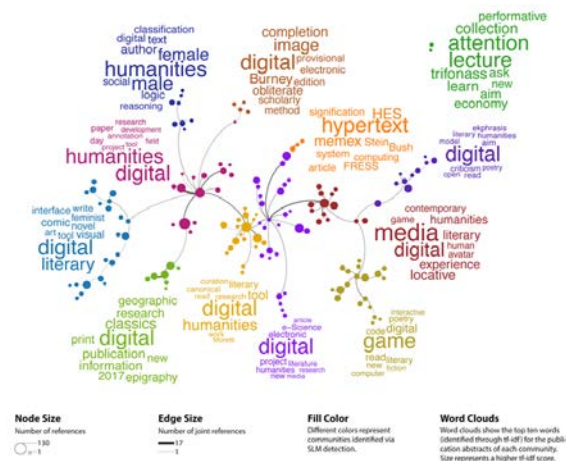
Fig. 3 Co-Author network by year

Figure 5 shows key papers in the backbone bibliographic coupling network, that is, the papers that link each of the communities in the giant component. The labels are shown in the same colors as the communities in Figure 4.

After we removed articles with zero references, the network contained 170 articles (out of 195), of which 23 are without a connection to other (i.e. they remained isolate). These 23 are not shown in the final visualization above, showing 147 articles and 145 connections. The bibliographic coupling network contains twelve communities, of which one consists of two articles not otherwise connected to the

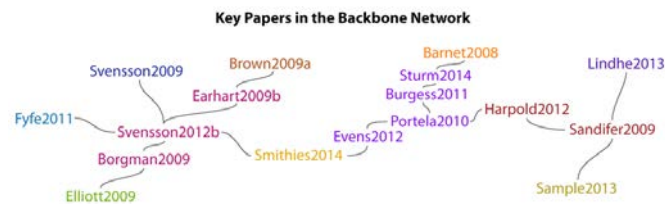
Fig. 4 Backbone bibliographic coupling network for DHQ.

## Backbone Bibliographic Coupling Network for Digital Humanities Quarterly: 2007-2014



major component (see dark green at the upper right). The other eleven communities are all connected in the large component and shown with their respective word clouds.

Fig. 5 Key papers in the backbone bibliographic coupling network.articles



There are a total of 4880 documents, including the 195 articles from DHQ itself. Together all the DHQ articles contain 5330 references. The highest cited document is Matthew Kirschenbaum's "*Mechanisms: New Media and the Forensic Imagination*" (2008), cited 15 times. The DHQ article with the most references is Christine Borgman's "*The Digital Future is Now: A Call to Action for the Humanities*" (2009), with 130 references.

### 3 DISCUSSION

#### 3.1 Co-author Network

The co-author network suggest that DHQ publications follow the patterns of the humanities community, with many single-authored papers (128 out of 195, 65.6%). Moreover, its origins are in North America, and three quarters of the authors are from either the US (58%) and Canada (17%). A distant third is the UK (9%), further demonstrating the Anglo-Saxon nature of DHQ.

The largest co-author network component consists of 43 authors; which is about 16% of all authors (276 authors in all) who contributed to DHQ during this period. The second largest co-author network component consist of 18 authors.

Canadian authors show the most collaborative behavior: the article with the most co-authors: "*Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET)*" has 14 co-authors. The most collaborative author in this period from Canada is Stan Ruecker; he co-authored 4 articles with 25 others.

There does not seem to be a growth of co-authorship after 2008. Overall, articles have on average a little under two authors per paper, and in 2012 a bit above two on average (2.18). When we remove all the single-authored papers, the average number of authors per article is above three, but there is no trend that this is growing with the years.

#### 3.2 Bibliographic coupling network with word clouds

From the word clouds we see that several communities explicitly discuss terms such as *digital* and *humanities* as well as *tool*, which is unsurprising. At the centre of the large component, the communities (magenta, yellow, purple) of articles are related to (textual) tools and discussing DH itself, with terms such as *curation*, *e-Science*, *project*, and *research*. The communities further to the left (light blue & dark blue) are related to textual analysis and tools, with terms such as *classification*, *author*, *write*, *annotation*, *interface*, and *literary*. The communities to the right however (dark purple, dark red, moss-

green) suggest articles related to artistic subjects, with terms such as *poetry*, *ekphrasis*, *games*, and *fiction*.

### 4 CONCLUSION

We return to the questions provided by the DHQ editors:

1. how citations reflect differences in academic culture at the institutional and geographic level
2. the changes to that culture over time.
3. correlations between article topics (reflected in keywords) and citation patterns.

With respect to the first question, we focus on the geographic level of academic culture. The co-author network shows that despite DH being a collaborative culture, over half of all publications are single authored, something demonstrated earlier for other journals<sup>9</sup>. Moreover, DH as represented by DHQ is largely an Anglo-Saxon North American undertaking. With respect to the second question; there is no visible trend regarding co-authorship between 2007-2014. However, authors from non-Anglo Saxon countries are emerging, showing DH is slowly becoming a more global phenomenon as also evidenced by the DH conferences<sup>10</sup>.

With respect to the third question, we find that the references present in the DHQ articles lead to a large number of communities. The boundaries are however diffuse, making it difficult to describe clear cut communities. However, from the word clouds we do see at least three different patterns emerge: 1) article related to tools and DH itself, 2) articles related to textual analysis with tools, and 3) articles related to artistic subjects.

While we have provided an exploration of the articles and authors within DHQ, additional insights may be learned from further analysis. In particular, interactive visualizations will provide the user with a more comprehensive understanding of the data. These may allow the user to explore communities via institution or discipline as well as country. In addition, we believe a properly de-biased authorial bibliographic coupling network may provide further insight into the academic cultures within DHQ. Lastly, our analysis focused on DHQ articles alone. Further analysis may allow us to explore the non-DHQ articles cited by DHQ papers.

In sum, we see DHQ fairly represents the heterogeneity of DH, critically examining DH itself and discussing computational analyses of research questions from different backgrounds. On the other hand, however, we see DHQ representing a somewhat homogeneous view of DH, with strong representation from Anglo-Saxon scholars and those from North America in particular. Here, DHQ can be challenged to provide a better representation of scholars from other backgrounds, as well as the 'big tent' of DH in general.

### ACKNOWLEDGMENTS

The authors wish to thank Professor Julia Flander, Professor Katy Börner, Dr. Andrea Scharnhorst, and the participants of Indiana University's Information Visualization MOOC for providing us valuable feedback during the process of the project work.

### REFERENCES

- [1] Svensson, Patrik. (2012) Beyond the big tent. *Debates in the Digital Humanities*, 36-49.
- [2] Knorr Cetina, K. (2007). *Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures*. The Blackwell

<sup>9</sup> <http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/10/joint-authorship-digital-humanities-collaboration>

<sup>10</sup> See <http://www.scottbot.net/HIAL/?p=41064>

Companion to the Sociology of Culture, 32(4), 361–375.  
doi:10.1002/9780470996744.ch5

- [3] Digital Humanities Quarterly (n.d.). About DHQ. Retrieved from <http://www.digitalhumanities.org/dhq/about/about.html>
- [4] Ortega, L., & Antell, K. (2006). Tracking Cross-Disciplinary Information Use by Author Affiliation: Demonstration of a Method. *College & Research Libraries*, 67(5), 446–462. Retrieved from <http://crl.acrl.org/content/67/5/446>.
- [5] Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, <https://sci2.cns.iu.edu>.
- [6] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." *ICWSM* 8 (2009): 361-362.
- [7] Jacomy, Mathieu, et al. "Forceatlas2, a continuous graph layout algorithm for handy network visualization." *Medialab center of research* 560 (2011).
- [8] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, 14(1), 10-25.
- [9] Schvaneveldt, R. W., D. W. Dearholt, and F. T. Durso. "Graph theoretic foundations of pathfinder networks." *Computers & mathematics with applications* 15.4 (1988): 337-345.
- [10] Waltman, Ludo, and Nees Jan van Eck. "A smart local moving algorithm for large-scale modularity-based community detection." *The European Physical Journal B* 86.11 (2013): 1-14.
- [11] Blázquez, M. (n.d). Frecuencias y pesos de los términos en un documento. Retrieved from: <http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com.es/2012/11/frecuencias-y-pesos-de-los-terminos-de.html>

### ***Appendix 6: Text and Slides for DH2015 Paper***

This appendix contains the text and slides for a paper on DHQ (mentioning but not focused primarily on the bibliographic project) presented at DH2015 in Australia: “Challenges of an XML-based Open-Access Journal: Digital Humanities Quarterly,” Julia Flanders, John Walsh, Wendell Piez, Melissa Terras. The text of this paper has been revised based on commentary and discussion in the conference session.

# Challenges of an XML-based Open-Access Journal: Digital Humanities Quarterly

Julia Flanders (Northeastern University)

John Walsh (Indiana University)

Wendell Piez (Piez Technologies)

Melissa Terras (University College London)

## 0. Introduction

Digital Humanities Quarterly was founded in 2005 as ADHO's first online open-access journal and published its first issue in 2007.

- In the ensuing ten years, the journal has been conducted as an ongoing experiment in standards-based journal publishing.
- In this paper we'd like to reflect on the results of that experiment to date, with emphasis on a few areas of particular challenge and research interest

During that period, other open-access journals in DH have also emerged, and if we look at them as a group we can see some differences of approach which reflect differences of goals and philosophy, and also the kinds of personnel and other resources they have available:

- Approach to the data: is the article data itself of interest as a potential future research asset? Does the community have a predilection towards a particular data format (e.g. TEI?)
- Approach to publication architecture: content management system (emphasizing configurability by novice administrators and design-oriented control over format) or data-driven approach (emphasizing consistent exploitation of the data with no design intervention except at the systemic level)
- Where does the mission reside? In the content or in the information system?

DHQ is perhaps an extreme example of a data-driven journal with an overwhelming interest in its own information systems, and this orientation arises in great part from the specific people to whom the journal's initial design and launch was entrusted: having a strong research interest in XML, in data curation, in future exploitation of the journal as a data source. This paper isn't intended as an exercise in evangelism or self-praise, but rather an exploration of what happens when we choose that set of parameters and follow their logic. The results thus far may help others working on developing open-access journals to situate their efforts within this same set of constraints.

## 1. Background and technical infrastructure

A few words about DHQ's fiscal and organizational arrangements may be useful here because they determine many of the strategic choices I'll be talking about. [slide]

- Funded jointly by ACH (which is the formal owner of the journal) and ADHO, each of which contributes \$6000 per year.
- As of 2014, also receives funding from Northeastern University for the managing editor positions, two graduate research assistants at 10 hours per week each during the academic year; Indiana University has also contributed staff time and services.
- Uses grant funding to support special projects (currently completing two small grant-funded projects which I'll describe a bit later)
- The journal is led by three general editors and a technical editor, together with an editorial team that has more specialized responsibilities
- The editor in chief oversees two managing editors and the overall workflow of submission, review, and production; and the technical editor oversees a Technical Assistant and the maintenance and development of the journal's technical systems (version control, servers, publication apparatus)

DHQ's technical design was constrained by a set of higher-level goals and needs.

- As an early open-access journal of digital humanities, an opportunity to participate in the curation of an important segment of the scholarly record in the field.
- Hence more than usually important that the article data be stored and curated in a manner that would maximize the potential for future reuse.
- In addition to mandating the use of open standards, this aim also strongly indicated that the data should be represented in a semantically rich format.
- Also anticipated a need for flexibility and the ability to experiment with both the underlying data and the publication interface, throughout the life of the journal, without constraint from the publication system.

All of these considerations moved the journal in the direction of XML (and eventually to TEI), which would give us the ability to represent any semantic features of the journal articles we might find necessary for either formatting or subsequent research. It would also permit us to design a journal publication system, using open-source components, that could be closely adapted to the DHQ data and that could evolve (at our own pace and based on our own agenda) to match any changes in requirements for the data. At the journal's founding, several alternative publishing platforms were proposed (including the Open Journal System), but none were XML-based and none offered the opportunity for open-ended experimentation that we needed.

DHQ's technical infrastructure is a standard XML publishing pipeline [slide] built using components that are familiar in the digital humanities:

- Cocoon: pipelining tool that manages user interactions
- XSLT to transform the XML
- CSS and a little JavaScript for formatting and behavior
- Eventually, an XML database to handle queries to bibliographic data

Workflow also uses generally available tools: [slide]

- Submissions are received and managed through OJS through the copyediting stage

- Final versions of articles are converted to basic TEI using OxGarage (<http://www.tei-c.org/oxgarage/>).
- Further encoding and metadata are added by hand
- Items from the articles' bibliographies are entered into a centralized bibliographic system that is also XML-based.
- All journal content is maintained under version control using Subversion. The journal's organizational information concerning volumes, issues, and tables of contents is represented in XML using a locally defined schema [slide].
- The journal uses Cocoon, an XML/XSLT pipelining tool, to process the XML components and generate the user interface.

Consider DHQ in relation to two other journals who are more or less in the same quadrant, Digital Medievalist (first issue in 2005) and jTEI (first issue in 2011), which have some similarities of approach to DHQ:

- Desire to keep data in semantically rich formats such as TEI
- Using open-source tools
- DM and jTEI both have developed publishing workflows based on their TEI data
- Neither journal is the sole proprietor of its own publishing system, so the evolution of their publishing platform is to some extent constrained by the goals of those platforms (driven by the entire community of users, not just that journal)
- Hence these journals benefit from advances by those communities but can't easily anticipate them or exercise a determining influence
- Whereas DHQ has the reverse problem: we are responsible for our own interface, so we are free to change it as much as we like, but we have to find the resources to do it ourselves.

## 2. DHQ's Evolving Data and Interface

As noted above, DHQ's approach to the representation of its article data has from the start been shaped by an emphasis on long-term data curation and a desire to accommodate experimentation, and our specific encoding practices have evolved significantly during the journal's lifetime.

- The first schema developed for the journal was deliberately homegrown, and was designed based on an initial informal survey of article submissions and articles published in other venues.
- Following this initial period of experimentation and bottom-up schema development, once the schema had settled into a somewhat stable form we expressed it as a TEI customization and did retrospective conversion on the existing data to bring it into conformance with the new schema.
- At several subsequent points significant new features have been added to the journal's encoding: for example, explicit representation of revision sites within articles (for authorial changes that go beyond simple correction of typographical errors), enhancements to the display of images through a gallery feature, and adaptation of the encoding of bibliographic data to a centralized bibliographic management system.

- At the beginning of our schema design process, we noted that at some point we might want to create a “crayon-box” schema whose elements would be deliberately designed to support author-specified semantics (slide), with the author also providing the display and behavioral logic, but we have not yet had a call for this approach and have not yet explored it in any practical detail.

These changes to the data have typically been driven by emerging functional requirements, such as the need to show where an article has been revised or the requirements of the special issue on comics as scholarship. However, they also respond to a broader set of requirements:

- that this data should represent the intellectual contours of scholarship rather than simply interface.
- For example, the encoding of revision notes retains the text of the original version, identifies the site of the revision, and supports an explanatory note by the author describing the reason for the revision. Although DHQ's current display uses this data in a simple manner to permit the reader to read the original or revised version, the data would support more advanced study of revision across the journal.
- Similarly, although our current display uses the encoding of quoted material and accompanying citations in very straightforward ways, the same data could readily be used to generate a visualization showing most commonly quoted passages, quotations that commonly occur in the same articles, and similar analyses of the research discourse. The underlying data and architecture lend themselves to incremental expansion.

### 3. Experimentation; Design vs. Data-driven approach

DHQ's data driven approach is rooted in caution and in motives of security, which are in a sense fundamentally conservative. Supporting the long-term preservability and intelligibility of our articles-as-data becomes much easier if that data is strongly convergent. Similarly, our task of publication is much easier and cheaper if our mechanisms of display are strongly determined by the data. However, one principle we articulated at the journal's launch was the idea that we wanted to support experimentation not just by ourselves but by authors, and we established a rationale for this experimentation that expressed its costs and risks and allocation of responsibility in terms of conceptual “zones”: [slide]

- Zone 1 is DHQ proper, using standard DHQ markup and display logic. Within Zone 1 we seek to provide an expanding set of functions that keep up with the most typical needs of DHQ authors. DHQ takes full and perpetual responsibility for maintaining Zone 1 articles in working order.
- Zone 2 is a space of collaborative experimentation between DHQ and the author, in which we can accommodate author-generated data and code under specified terms:
  - it must meet certain standards of curatability: using open standards and formats, using tools and languages that make sense for DHQ to maintain expertise in
  - it must conform to good practice (documentation, commented code) so that the code itself can be considered a publication, not just an instrument of getting something done



- it must include an XML fall-back description so that if the experimental version breaks, readers can still find an intelligible account of it, and also to provide some kind of basic operation and discoverability within DHQ's standard search mechanisms
- DHQ takes a more cautious form of responsibility for articles in Zone 2: we'll curate the data and we'll do our best to keep the code working, but we can't guarantee that we'll support all of its dependencies in the future since we can't be sure our resources will support that level of effort
- Zone 3 is a space of authorial autonomy, with many fewer constraints on the author and greatly diminished responsibility on DHQ's part:
  - The code needs to be something that can actually run on DHQ servers without risk, or else the author can host it on his/her server
  - The code needs to conform to good practice (documentation and commenting)
  - There needs to be an XML fall-back description, which is even more important in this case because the likelihood of fragility is so much greater

So it's interesting to consider at this point what forms that experimentation might take: how do authors actually want to experiment, and how far are we actually prepared to go to support them?

At a very simple level:

- we can observe that authors do want control over formatting, and this gives us a window into what "authoring" in the digital medium entails.
- the most common kinds of requests or push-back we get from authors have to do with layout: the formatting of tables, the placement and sizing of images, the fine-tuning of epigraphs and code samples.
- Note that these are all components with a strong visual component to their rhetoric; unlike paragraphs and notes and block quotations and citations, in which the strength of the semantic signal is so strong that we receive their full informational payload regardless of how they are formatted, these visual features have the potential to mean differently or less successfully if they look different.
- These are also all features for which it would be comparatively easy for DHQ to provide finer mechanisms of control simply by making our own stylesheets more elaborate (asking them to handle more article-specific renditional information, and taking the trouble to work out the potential collisions and tricky cases): so the chief limiter here is cost.

At a more advanced level, authors might experiment by proposing new semantic features. The actual examples so far have been features that are recognizable but that we just hadn't anticipated and hadn't developed any specific encoding for:

- Timelines
- Annotated bibliography
- Survey data
- Oral history interviews

We have the choice here of representing these as if they were more generic features we already support (an oral history interview is a dramatic dialogue; a timeline is a kind of list), or of treating them as semantically distinct. The most compelling motivation for the latter approach would be the possibility of strengthening our support for the study of discourse, which would entail having a larger set of instances: so here, the role of the initial experiment is to bring a given feature to our notice but the work of actually supporting it is only warranted if it's a feature other people want as well.

We have also had a few examples of genuinely experimental writing in which the author was deliberately departing from the genre of the scholarly article. (Slides: Trettien).

- The question we have to ask here is: are these experiments in semantics or in design? We've seen that a journal like DHQ can in principle accommodate authorial control over display (at a cost), and as we noted earlier, we have at least theoretically entertained the idea of allowing authorially specified semantics through a specialized schema. The question is, which are these experimental authors asking for?

If we examine these cases more closely, a few points are worth noting:

- The experimental cases so far have been expressed as JavaScript and HTML, and their rhetorical innovation takes the form of textual behaviors: responsiveness to reader actions (mouseover, clicking) in the form of navigation and motion, the text moving or changing form.
- In other words, they emphasize effects which are significant precisely because they depart from display norms; the Trettien piece plays on our expectations of textual fixity and accuracy, and the Bianco piece thwarts our expectations about reading one thing at a time
- However, they don't seem to introduce a new semantics, a new rhetorical feature that they could usefully declare through their encoding: the innovation lies in what they *do* rather than in what they *are*; it lies precisely in how the reader will experience the surface of the text rather than in what the reader might do if he/she could get at the underlying data and work directly with that. Giving the reader access to "the data" would give the reader nothing at all of what is actually going on in these pieces.

So far, we have not had any proposed experiments that work in the other direction. What would they look like?

- An article that does exactly what Trettien did, but using XML rather than HTML as the source data
- An article that is mostly structured data (e.g. data from a survey) with XSLT that presents it to the reader for inspection and manipulation (sorting, filtering)
- A special issue that uses a TEI customization and for which the guest editors have developed XSLT and CSS that exploits the articles' markup

The best way for us to pursue this kind of experimentation would be to invite proposals, perhaps structured around a grant proposal to provide some support for stylesheet development. (Consider this an informal invitation!)

## 4. Next Steps

DHQ has several developmental projects under way:

- With generous support from a grant organized by Marco Büchler from the University of Leipzig, we are implementing an OAI-PMH server for DHQ through which we can better expose the journal's metadata
- [slide] We have just completed an NEH DH startup grant which funded the development of a centralized bibliography for DHQ: important improvement for DHQ's production processes, but also opens up some exciting potential for citation analysis and data visualization; we'll be publishing an article about this in the coming months
- We are also in the planning stages of a project to explore internationalization of the journal through a series of special issues dedicated to individual languages. This will involve some further work on the schema and interface, and also changes to the workflow to accommodate a multilingual review process. We will be working within our existing constraints of finances and personnel so we'll need to proceed deliberately, but we're excited to be undertaking this step.

# Challenges of an XML-based Open-Access Journal: Digital Humanities Quarterly

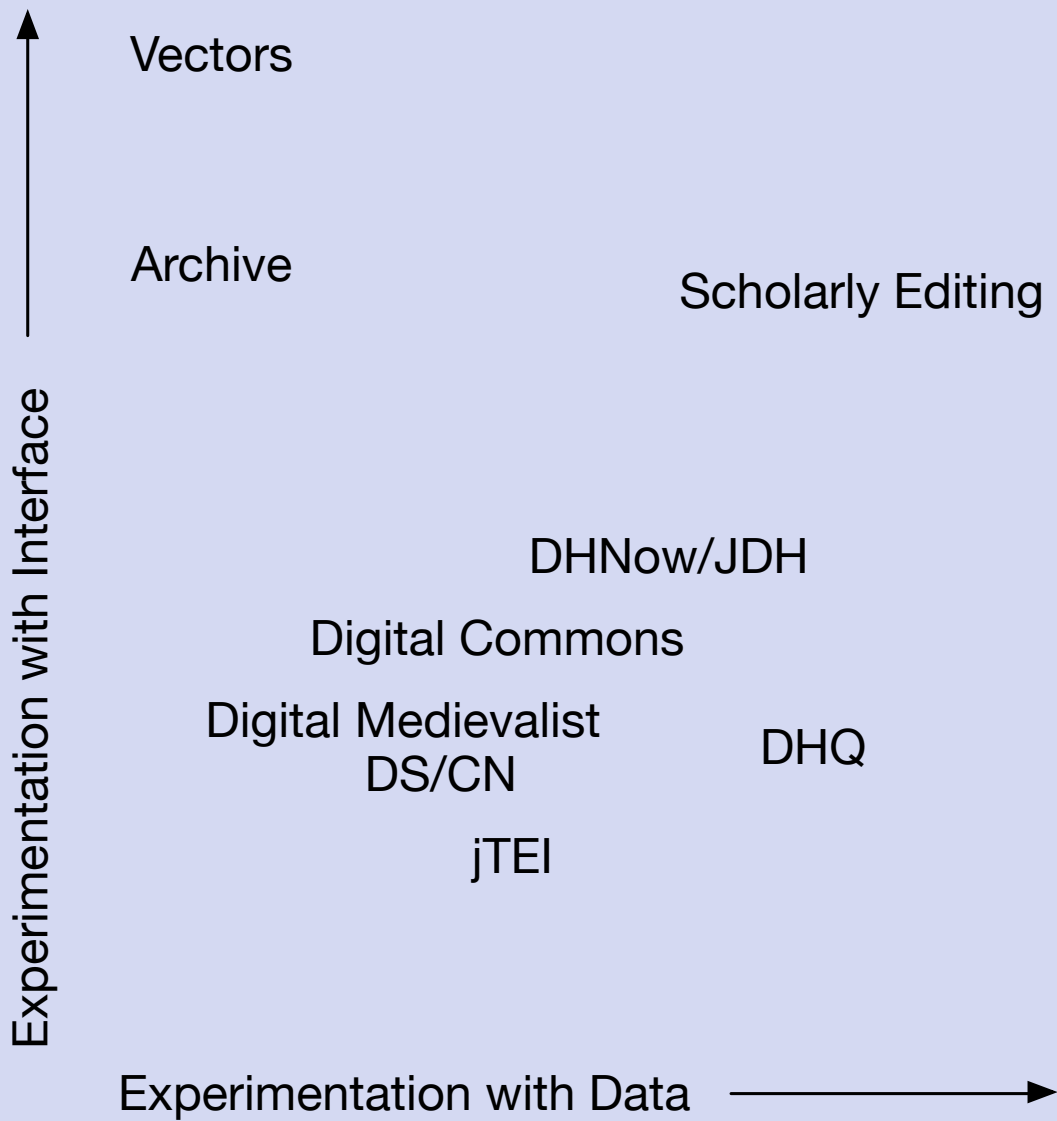
Julia Flanders, Northeastern University

John Walsh, Indiana University

Wendell Piez, Piez Consulting

Melissa Terras, University College London





# Background on DHQ

- Founded in 2005, first issue in 2007
- Jointly funded by ACH and ADHO
- Hosted and supported at Northeastern University and Indiana University
- Grant-funded special projects

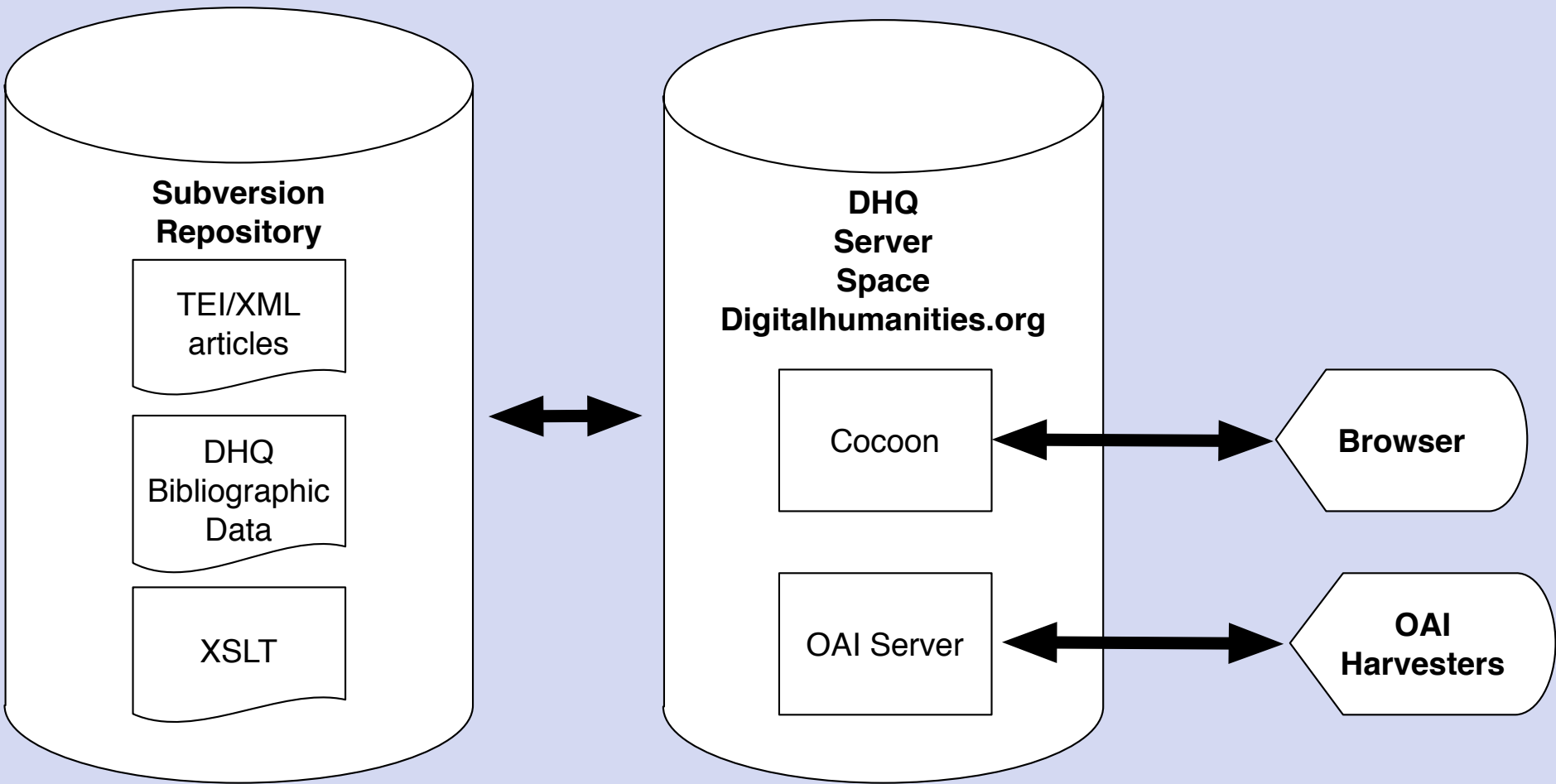


# Staff and organization

- General Editors: Julia Flanders, Wendell Piez, Melissa Terras
- Technical Editor: John Walsh
- Managing editors: Elizabeth Hopwood, Duyen Nguyen, Jonathan Fitzgerald
- Technical assistant (currently vacant)
- Editorial team: Stéfan Sinclair, Adriaan van der Weel, Alex Gil, Michelle Dalmau, Jessica Pressman, Geoffrey Rockwell, Sarah Buchanan
- Special teams players: Jeremy Boggs
- Abundant excellent peer reviewers

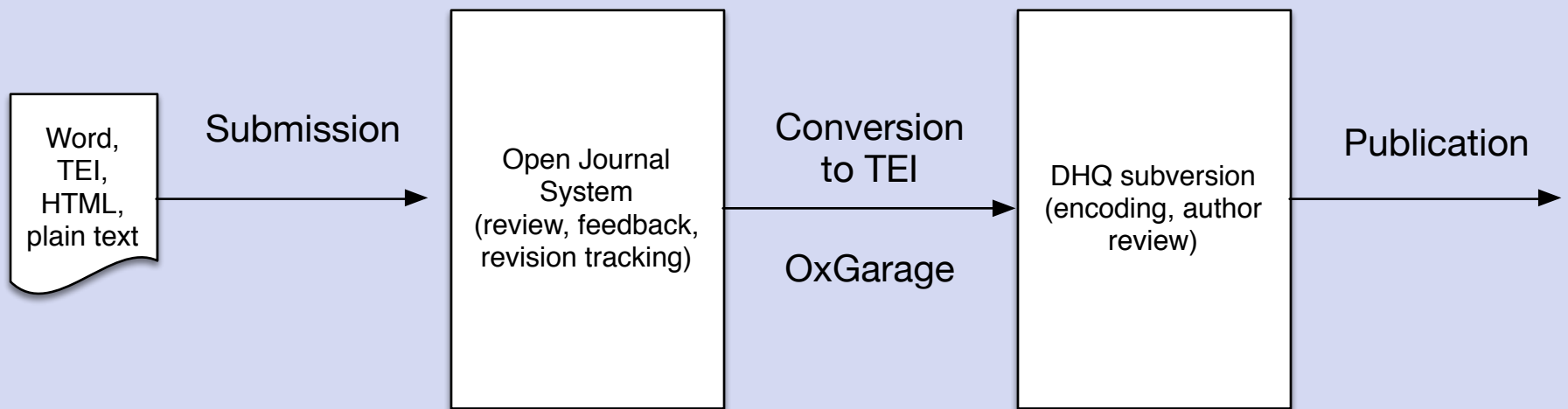






**Northeastern University**

*NULab for Texts, Maps, and Networks*



```
<journal vol="9" issue="1" current="true">
  <title>2015</title>
  <list id="editorials">
    <item id="000206"/>
  </list>
  <list id="articles">
    <item id="000205"/>
    <item id="000203"/>
    <item id="000199"/>
    <item id="000198"/>
    <item id="000204"/>
  </list>
  <list id="reviews">
    <item id="000207"/>
  </list>
</journal>
```



<div type="" subtype="" rend="">

<head>An Experiment in XML</head>

<ab type="" subtype="" rend="">Experimental text block with behaviors controlled by stylesheets and the possibility of <seg type="" subtype="" rend="">inline elements</seg> whose formatting and behavior are also controlled by stylesheets.</ab>

<user:myElement>Namespaces could also be used to include user-defined elements (or elements from other established XML languages) with specified semantics.</user:myElement>

</div>



### Revision Note

This article has been revised since its original publication. A response solicited by the author from Matthew Kirschenbaum has been added as a [footnote](#). The [previous version of the article](#) will remain available.

<dhq:revisionNote previous="000128\_01" when="2013-04-17">

This article has been revised since its original publication. A response solicited by the author from Matthew Kirschenbaum has been added as a  
<ref target="#revItem\_01">footnote</ref>.

</dhq:revisionNote>



<b>Zone</b>	<b>Features</b>	<b>Curation</b>
Zone 1	DHQ markup and stylesheets	DHQ in perpetuity
Zone 2	Author-supplied code, constrained by DHQ support capabilities; Zone 1 fallback required	DHQ good faith curation
Zone 3	Author-supplied code, constrained by good practice guidelines; Zone 1 fallback required	No DHQ responsibility



First, though, one must recognize Milton's tract as presenting a theory of mediation. Until recently ([[Dobranski 1999](#)]; [[Fulton 2010](#)]; [[Norbrook 1994](#)]; [[Rovira 2005](#)]), interpretive readings of *Areopagitica* have been locked in a debate about the degree of liberalism or conservatism Milton advocates in the censorship of books, with Stanley Fish's reader-response-oriented reading dominating the critical discourse ([[Fish 2001](#)]; on this debate, see [[Kolbrener 1997](#), 1-7]). While much of this work has constructively problematized the tract's politics, it is important to emphasize that, for Milton, books (including his own) are not passive containers for disseminating political ideas but, as he writes in a well-known passage,

doe contain a potencie of life in them to be as active as that soule was whose progeny they are; nay they do preserve as in a violll the purest efficacie and extraction of that living intellect that bred them. I know they are as lively, and as vigorously productive, as those fabulous Dragons teeth; and being sown up and down, may chance to spring up armed men. [[Milton 1644](#), 4]



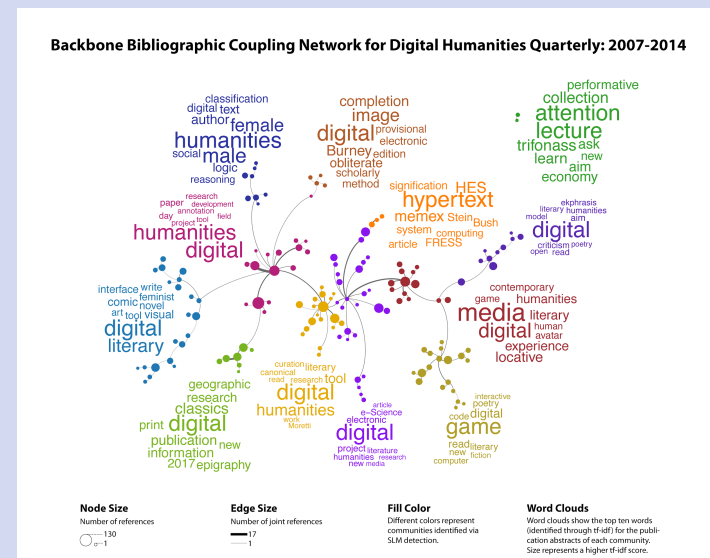
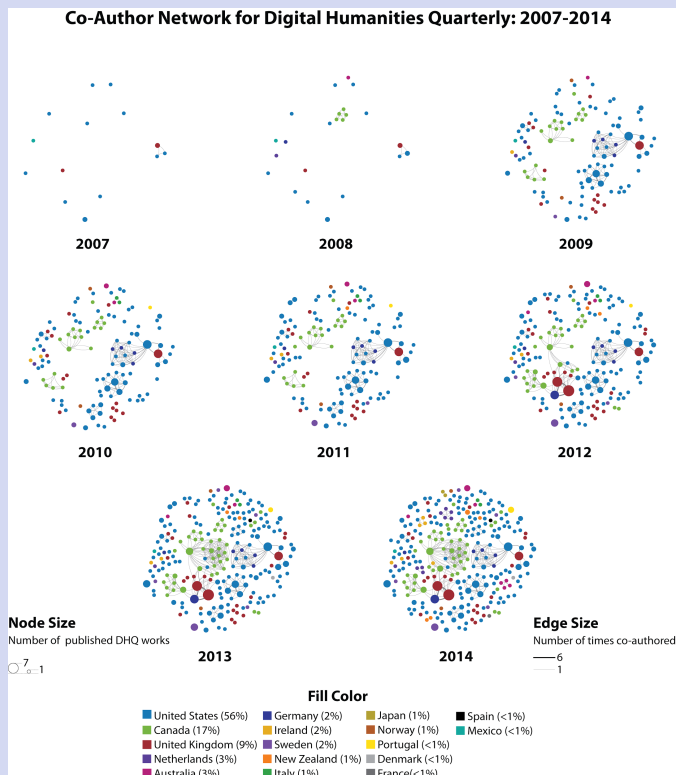


doe contain **a~em** to be as active as that soule was whose progeny they are; nay they do preserve as in a viol the purest efficacie and extraction of that living intellect that bred them. I know they are as lively, and as vigorously productive, as those fabulous Dragons teeth; and being sown up and down, may chance to spring up armed men. [[Milton 1644](#), 4]

doe contain a potencie of life in them to be as active as that soule was whose progeny they are; nay ~~they do~~ preserve as in a viol the purest efficacie and extraction of that living intellect that bred them. I know they are as lively, and as vigorously productive, as those fabulous Dragons teeth; and being sown up and down, may chance to spring up armed men. [Milton 1644, 4]

dead things, but doe contain a potencie of life in them  
to be as active as that soule was whose progeny they are;  
nay, they do preserve as in a viol the purest efficacie  
30 and extraction of that living intellect that bred them.  
I know they are as lively, and as vigorously productive,  
as those fabulous Dragons teeth; and being sown up  
and down, may chance to spring up armed men. And





“Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly,” Dulce Maria de la Cruz, Jake Kaupp, Max Kemman, Kristin Lewis, and Teh-Hen Yu. Final project submitted for Information Visualization MOOC, Indiana University, May 2015.



Thank you!

Julia Flanders @julia\_flanders

John Walsh

Wendell Piez

Melissa Terras @melissaterras

